U.PORTO

MESTRADO
MULTIMÉDIA - ESPECIALIZAÇÃO EM TECNOLOGIAS

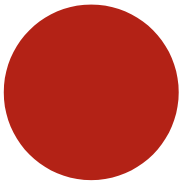
DESIGNING AN ALGORITHMIC NEWS APPLICATION FOR USER CONTROL

Tiago Nuno Mesquita Folgado Leitão Devezas

M
2017

FACULDADES PARTICIPANTES:

FACULDADE DE ENGENHARIA
FACULDADE DE BELAS ARTES
FACULDADE DE CIÊNCIAS
FACULDADE DE ECONOMIA
FACULDADE DE LETRAS



U.PORTO

Designing an Algorithmic News Application for User Control

Tiago Devezas

Master in Multimedia, University of Porto

Supervisor: Sérgio Sobral Nunes (Assistant Professor)

July 19, 2017

Designing an Algorithmic News Application for User Control

Tiago Devezas

Master in Multimedia, University of Porto

Approved in oral examination by the committee:

Chair: Rui Luís Nogueira Penha

External Examiner: António Maria Salvado Coxito Granado

Supervisor: Sérgio Sobral Nunes

Abstract

The torrential stream of information published online every day has fostered the rise of technological solutions aiming to combat information overload. News aggregators, applications that collect news from multiple publishers and present them in a condensed form in a single place, are one such solution. In order to increase efficiency, many aggregators now employ personalization and curation algorithms, computer programs that decide what to present based on a set of rules and criteria. But this approach can entail new problems. Algorithms can be biased, lock users inside "filter bubbles", and help disseminate false news. This power to shape public opinion is often left unchecked. Information about the algorithm's inner workings isn't provided to users, who sometimes aren't even aware that their news feed is managed by a computer program.

This study aims to identify the importance of a set of criteria that a sample of potential users would like to manipulate in an algorithmically generated news aggregator, assess the importance of providing information about the algorithm, and determine the desired degree of control over the system. The goal is to then apply the most relevant findings to a functional prototype in line with the users' requirements. In order to achieve these objectives, a quantitative non-experimental research design was employed. A questionnaire was developed and sent via email to the University of Porto community. Four hundred and thirty two (432) participants, mainly comprising students and professors, filled the questionnaire. Data was collected about demographics, news consumption habits, and the importance level of multiple criteria regarding news content, diversity of perspectives, news sources, information about the algorithm, and the degree of personalization.

Results indicate that the most valued news content criteria were "Timeliness", "Good news", "Relevance", "Conflict", "Magnitude", "Serendipity", "Audio-visual", and "Proximity". Regarding the diversity of perspectives, participants attributed higher importance to news that present points of view diverse from their own than to content aligned with their views. Professionalism and journalistic prestige were rated as the most important factors concerning the sources. Information about the algorithm was evaluated as being of moderate importance. As for personalization, participants indicated that the most important criterion was the ability to explicitly configure the system according to their own options.

This study provides some insights about the importance of several criteria for manipulating an algorithmic news application from the users' perspective. Some of the findings were transposed to a functional prototype that will be further developed in the future. Evaluating the system with users is also planned as future work. The usage of a non-probability sample limits the generalization of the findings, but we believe they can be a valuable reference for implementers of algorithmic news applications concerned with user control and transparency.

Resumo

A torrente de informação publicada online todos os dias fomentou o aparecimento de soluções tecnológicas que visam combater a sobrecarga de informação. Os agregadores de notícias, aplicações que recolhem notícias de várias fontes e apresentam-nas de forma condensada num único local, são uma destas soluções. Para aumentar a eficiência, muitos agregadores recorrem atualmente a algoritmos de personalização e curadoria, programas de computador que decidem o que apresentar com base num conjunto de regras e critérios. Mas esta abordagem pode criar novos problemas. Os algoritmos podem conter enviesamentos, fechar os utilizadores dentro de "bolhas de filtro", e ajudar a disseminar notícias falsas. Este poder de moldar a opinião pública é muitas vezes deixado sem controlo. Informação sobre o funcionamento interno do algoritmo não é fornecida aos utilizadores, que por vezes nem estão conscientes de que a sua lista de notícias é gerida por um programa de computador.

Este estudo visa identificar a importância de um conjunto de critérios que uma amostra de potenciais utilizadores gostaria de manipular num agregador de notícias gerado por um algoritmo, aferir a importância de fornecer informação sobre o algoritmo, e determinar qual o grau de controlo sobre o sistema que os utilizadores pretendem. O objetivo é depois aplicar as conclusões mais relevantes a um protótipo funcional alinhado com os requisitos dos utilizadores. De forma a atingir estes objetivos, foi desenhado um estudo quantitativo não-experimental. Desenvolveu-se um questionário, que foi enviado por email à comunidade da Universidade do Porto. Quatrocentos e trinta e dois (432) participantes, maioritariamente estudantes e docentes, preencheram o questionário. Foram recolhidos dados sobre as características demográficas, hábitos de consumo de notícias, e o nível de importância de vários critérios relacionados com o conteúdo das notícias, diversidade de perspetivas, fontes noticiosas, informação sobre o algoritmo, e o grau de personalização.

Os resultados indicam que os critérios relacionados com o conteúdo das notícias mais valorizados foram "Atualidade", "Boas notícias", "Relevância", "Conflito", "Magnitude", "Serendipidade", "Áudio-visual", e "Proximidade". Relativamente à diversidade de perspetivas, os participantes atribuíram maior importância às notícias que apresentam pontos de vista diferentes dos seus do que a conteúdos alinhados com os seus pontos de vista. Profissionalismo e prestígio jornalístico foram avaliados como os fatores mais importantes no que se refere às fontes. Informação sobre o algoritmo foi avaliada como moderadamente importante. Quanto à personalização, os participantes indicaram que o critério mais importante é a possibilidade de configurar explicitamente o sistema de acordo com as suas próprias opções.

Este estudo proporciona algum conhecimento sobre a importância de vários critérios para manipular uma aplicação algorítmica de notícias segundo a perspetiva dos utilizadores. Algumas das descobertas foram transpostas para um protótipo funcional que continuará a ser desenvolvido no futuro. A avaliação do sistema com utilizadores está também planeada como trabalho futuro. A utilização de uma amostra não-probabilística limita a generalização dos resultados, mas acreditamos que estes podem ser uma referência valiosa para implementadores de aplicações algorítmicas de notícias preocupados com o controlo dos utilizadores e a transparência.

Acknowledgements

I dedicate this achievement to my dear departed mother. Wherever she may be, I am sure she is proud. I would like to thank my supervisor, Professor Sérgio Nunes, for his efficient guidance, patience, and overall niceness. Also, a word of appreciation to my colleagues from the InfoLab, particularly José Luís Devezas, with whom I have very interesting and instructive discussions. Finally, a word of affection and gratitude to my family and loved ones.

Tiago Devezas

"Of course you can't 'trust' what people tell you on the web anymore than you can 'trust' what people tell you on megaphones, postcards or in restaurants. Working out the social politics of who you can trust and why is, quite literally, what a very large part of our brain has evolved to do."

Douglas Adams

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The web is increasingly the medium of choice for news consumption. According to a 2016 survey, 38% of United States of America (USA) adults often get news from digital sources, such as news websites, apps and social networking sites (Pew Research Center, 2016b). While television still dominates, with 57% of USA adults getting news from it, web sources have already surpassed radio (25%) and print newspapers (20%). In addition to consumption, the web has also fostered a shift in content creation. Advances in software have significantly lowered the barriers to publish and share information online, allowing former consumers to become producers (the "prosumers"). The term "Web 2.0" (O'Reilly, 2005) is commonly employed to refer to the services and tools behind this transition.

While it is undeniable that the ability to quickly access and contribute to such a rich and diverse collection of knowledge is beneficial, it is also true that it can entail adverse effects. Two issues are commonly mentioned. While the rate at which new information is produced is virtually unlimited, the resource for which it competes, attention, is not. This mismatch between the enormous volume of information available and the limited cognitive ability to process it, known as information overload, can impair comprehension and decision making. A second issue is related with the information's credibility. Since there aren't any mandatory credibility checks to publish information online, the onus of assessing it lies with the consumer, who therefore incurs in additional cognitive load.

Algorithmic content personalization and curation is one strategy that can help to reduce information overload and increase the user experience (Rader, 2017) by tailoring the information displayed based on the system's assertion of what the user wants (Bozdag, 2015). However, it has been shown that this approach can introduce a new set of problems. Algorithms do not operate objectively and impartially, there are decisions, choices, and influences, embedded into them by humans. They can be biased (Friedman and Nissenbaum, 1996). Despite potentially reducing information overload by only showing users information deemed relevant, content personalization and curation algorithms can create "echo chambers" and "filter bubbles". If the information presented is mostly about issues with which users already agree, limiting their exposure to different perspectives can amplify biases and isolate them in their own ideological and cultural bubbles.

Additionally, particularly in systems without any human intervention, the issue of information credibility still applies. One such example is Facebook's News Feed, which uses hundreds of signals to determine what stories to display to users, including the number of shares, likes, and comments. This means that highly popular and shared pieces of content, despite being false, can be rewarded by the system and displayed in more feeds. This prominence can then lead to more shares, likes, and comments, therefore creating a self-reinforcing cycle. The rampant dissemination of fake, but highly shared, news on Facebook related with the 2016 USA presidential election (Silverman, 2016), coupled with the fact that 44% of USA adults get news on the site (Pew Research Center, 2016a), led to discussions about the potential impact of false news on the electoral outcome.

It has since been shown that most of the fake news stories disseminated on social media favoured the winning candidate, Donald Trump, over the runner-up, Hillary Clinton (Allcott and Gentzkow, 2017). Some commentators have expressed their belief that fake news shared on social media helped elect Donald Trump (Parkinson, 2016). However, recent research has indicated this is unlikely (Allcott and Gentzkow, 2017). After being criticized for letting fake news run rampant on its platform, Facebook has recently informed that it's working on implementing mechanisms to curb the spread of false stories (Facebook, 2016b). Even though in this case initial studies indicate that the election's outcome wasn't determined by the spread of misinformation, studies have shown that algorithmic manipulation can indeed shift voting preferences. Epstein and Robertson (2015) conducted a large scale study which estimated that the manipulation of a search engine's ranking algorithm could change the outcome of more than 25% elections worldwide.

Despite the increasingly important role played by algorithms in our society, their presence and operation is still mostly opaque. It has been shown that many Facebook users are unaware that their news feed is generated by algorithms and that, due to this unfamiliarity with the system's operation, they "make inferences about their relationships, wrongly attributing the composition of their feeds to the habits or intent of their friends and family" (Eslami et al., 2015).

1.1 Motivation and Goals

While helpful to tackle information overload, the combination of news aggregation and personalization and curation algorithms can entail nefarious outcomes. These include the spread of misinformation, locking users in cultural and ideological bubbles or amplifying existing ones, making wrong inferences due to a lack of awareness of the algorithm's presence, or even possible interferences in democratic processes. These potentially far-reaching effects, in conjunction with the fact that many systems which employ algorithms do not reveal their influence or operation, have led to a growing discussion, inside and outside of the academia, about how to achieve algorithmic transparency and accountability.

However, the discussion about algorithmic transparency in news media is still in its infancy. Normative approaches are lacking, and concerns about possible negative impacts on the users' experience have been mentioned. Some potential strategies have been discussed in the literature.

These include disclosing the algorithms' existence and influence, providing transparent information about how they operate, and let users manipulate the system, not the other way around.

This study aims to contribute to this field by achieving two goals. This first is to identify a set of criteria that users would like to manipulate in an algorithmic news aggregator, assess the importance of providing transparency information about the algorithm's presence and operation, and determine how much control they desire to have over the system. To do so, we'll address the following quantitative research question and sub-questions:

- In the context of an algorithmic news application, which criteria do users rate as most important to manipulate?
  - Regarding the news content
  - Regarding the diversity of perspectives in the news
  - Regarding the publishers (sources)
  - Regarding the information about the algorithm's presence and inner workings
  - Regarding the degree of control (personalization) over the system

The second objective is to analyse the quantitative data resulting from answering the questions above and employ those findings, along with insights from previous works, to develop a functional prototype of an algorithmic news aggregator. The system will aim to address the aforementioned issues of information overload, user control, and algorithmic transparency, while providing a pleasant experience based on the requirements expressed by the participants. Therefore, we intend to respond the following qualitative research question:

- How can the findings from this study be transposed to an algorithmic news aggregator concerned with user control and transparency?

1.2 Document Structure

This document is divided into six chapters, including this one, Chapter 1. Chapter 2 reviews the state of the art and previous works related to information aggregation, algorithms, algorithmic transparency, and their intersection with the field of news media. Chapter 3 reports the design employed in this study, the sample and its characterization, and the instruments and techniques used to gather and analyse data. In Chapter 4, we present and discuss the results from the statistical analysis of the data collected. Chapter 5 describes the implementation of the functional prototype and how the findings from this study were used to inform its development. Finally, in Chapter 6, we present the conclusions, main outcomes, and limitations of this study, discuss the contributions, and identify future lines of work.

Chapter 2

State of the Art

This chapter presents the state of the art and previous works related to news aggregators and algorithms and their interplay with news media. In addition to the definitions and typologies of the concepts under investigation, we identify the pros and cons that these systems and technologies entail and survey potential solutions to the issues that can arise. We also present findings from some user studies related with the investigated topics that can potentially guide us during the implementation of an algorithmic news aggregator, one of the expected outcomes of this work.

2.1 News Aggregators

Broadly, a news aggregator can be defined as "a website that takes information from multiple sources and displays it in a single place." (Isbell, 2010). Coddington (2015) builds upon this definition, adding that news aggregators collect information already published and present it in an abbreviated form (through headlines, summaries or excerpts). Thus, news aggregation can be defined as "taking information from multiple published sources and displaying it in an abbreviated form within a single place." (Coddington, 2015). In these definitions, the word "sources" refers to the content-producing organizations, not to the traditional journalistic sources often found in the literature, i.e., the people or entities who provide information to journalists.

Other definitions, despite classifying news aggregators similarly as the ones above, distinguish them from outlets that produce original material (Chowdhury and Landoni, 2006; Stanyer, 2009). The lack of original reporting is at the core of the criticism presented against news aggregators by their most vocal detractors, the so-called legacy news organizations, such as newspapers, television and radio networks (Anderson, 2013). The audiences' growing shift to the internet has been accompanied by a decline of the traditional media's profits. This has lead many journalists and news executives to establish a causal nexus between the two events and accuse news aggregators of diverting traffic using stolen content (Isbell, 2010).

There are, however, those who disagree with this stance. Journalist and professor Jeff Jarvis (as cited in Anderson, 2013) considers that aggregators bring value to the table by creating an audience for the content produced by news organizations. Anderson (2013) also frames aggregation as a

practice that can add value to existing content. Through the creation of bundles of linked articles ranked according to criteria such as importance, popularity, and newsworthiness, aggregators can provide structure, order, and context, to multiple independently produced news stories. Anderson (2013) notes that many of the skills required for traditional journalism, such as writing, news sense, and visual presentation, are also paramount in news aggregation. It should be noted, that in the context of Anderson's work, the news aggregators are not the websites *per se*, but the people who manually collect information from multiple sources and use it to produce new pieces of content.

The distinction between automated and manual collection processes is an important one, because it highlights the existence of many types of news aggregators. For instance, the main target of the traditional media companies criticisms are the automated aggregators owned by internet companies, such as the Google News[1] service (Anderson, 2013). This disparity had led to some attempts to categorize the different types of news aggregators.

2.1.1 Types of News Aggregators

According to Isbell (2010), news aggregators can be grouped into four categories: Feed Aggregators, Specialty Aggregators, User-Curated Aggregators, and Blog Aggregators.

Feed Aggregators, such as Google News[1], collect material from multiple websites and organize it into feeds, displaying the stories' headlines, excerpts and links to the full content on the originating source.

Specialty Aggregators also collect information from multiple sources but restrict their scope to a particular topic. An example is the technology news aggregator Techmeme[2].

User-Curated Aggregators, such as Reddit[3] and Digg[4], are websites where the content is submitted by the users. Content is usually picked from a more eclectic set of sources than most news aggregators, including blog posts, videos or pictures.

Blog Aggregators are blogs that produce content based on original material collected from other sources. The Huffington Post[5] and Gawker[6] (recently discontinued) are two examples of such aggregators.

Despite providing an useful baseline for comparison between distinct types of aggregators, some limitations of this model have been pointed. Coddington (2015) argues that this typology has issues of classification (in the case of specialty and blog aggregators) and mutual exclusivity (an aggregator from one type can pertain to any of the others).

[1] http://news.google.com
[2] http://techmeme.com
[3] http://reddit.com
[4] http://digg.com
[5] http://huffingtonpost.com
[6] http://gawker.com/

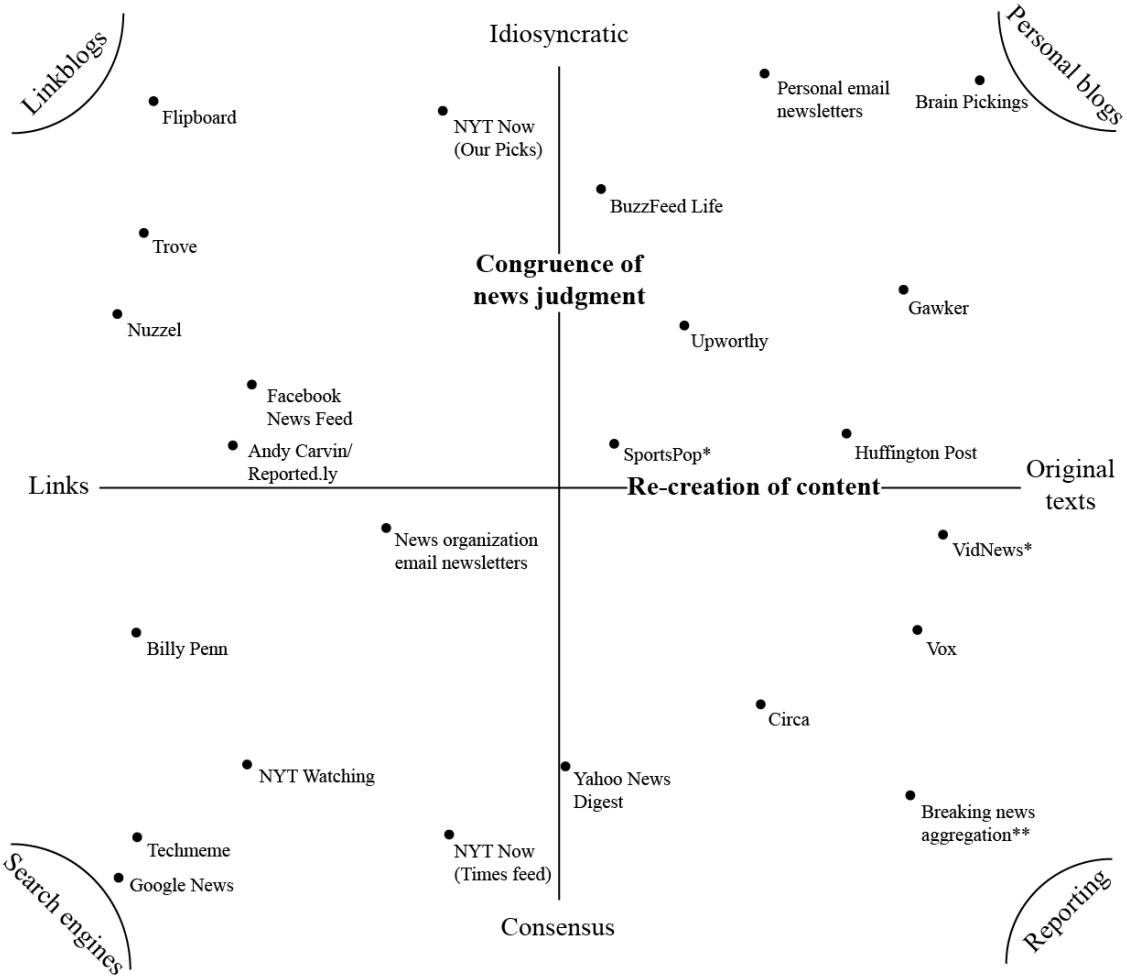


Figure 2.1: Typology of News Aggregators (Coddington, 2015)

The author proposes a different typology involving two dimensions: a horizontal one that measures the degree to which the content is recreated, and a vertical dimension representing the degree to which the content offered adheres to the professional consensus of newsworthiness and authority (Figure 2.1). Automated aggregators, which provide minimal or no recreation of content fall on one end of the first dimension, while aggregators which use information gathered from other sources as a starting point for new content are located near the other pole. As for the second dimension, on one end are the aggregators aiming to provide the most newsworthy stories from prominent sources, and on the opposite pole are the ones focusing more on eclectic and niche content.

2.1.2 User Experience and Usability of News Aggregators

Research about the user experience and usability of news aggregation websites is, to our knowledge, practically non-existent. One exception is the study conducted by Chowdhury and Landoni (2006). In order to inform the development of a novel news aggregation service, the

authors sought to identify what users expect and value in such platforms. Five aggregators were evaluated — Headlinespot[7], TVEyes[8], Newsburst (discontinued), Google News[1], and Awasu[9] — through an online questionnaire (45 participants) and interviews (10 participants). The questionnaire was divided in two parts: the first to assess user expectations of a news aggregators, and the second to collect information about the user experience with the evaluated services. Results reveal the most highly rated features in news aggregators were:

- Advanced search functionalities (considered essential by 80% of users)
- User-friendly interface (78%)
- High quality and reputable sources (76%)
- Search the past for related stories (71%)
- Browsing functionalities (69%)
- Presenting stories in chronological order (65%)
- Personalization (62%)

Other popular choices (with around 50% of users) were summarization, geographical personalization, multimedia features, ability to follow stories, and alert service functionalities. Despite the limitations pointed by the authors — limited sample size and a reasonably homogeneous user group —, these findings provide some guiding principles for the development of a news aggregation system.

2.2 Algorithms and News

An algorithm is a "systematic procedure that produces — in a finite number of steps — the answer to a question or the solution of a problem" (Britannica). Due to the increasingly relevant role algorithms play in our lives, they have been called the "new power brokers in society" (Diakopoulos, 2014). This is not an overstatement. For instance, in a study consisting of five experiments conducted with more than 4500 participants in two countries, Epstein and Robertson (2015) indicate that the manipulation of a search engine's ranking algorithm could shift the voting preferences of undecided voters by 20% or more (up to 80% in some demographics), with almost no visible traces of manipulation. As many elections are won by close margins, the authors estimated that the outcome of more than 25% elections worldwide could be changed through search ranking manipulation. The implications of this effect led one of the authors to affirm it's a "serious threat to the democratic system of government." (Epstein, 2015).

[7] http://www.headlinespot.com/
[8] http://www.tveyes.com/
[9] http://www.awasu.com/

In the field of news media, algorithmic influence is mainly exerted through content curation and personalization, i.e., the tailoring of information based on the algorithm's assertion of what the user needs, wants and his/her social connections (Bozdag, 2015). By connecting users with the information the system presumes they want, these algorithms aim to reduce information overload and improve the user experience (Rader, 2017).

The systematic and automated aspects of algorithms might lead to the belief that they remove human bias and operate objectively and impartially. However, contrary to this assertion, algorithms are not free of bias (Friedman and Nissenbaum, 1996). Algorithms are programmed by people, therefore imbued with human influence, such as criteria choices, training data, semantics and interpretation (Diakopoulos, 2014).

Another issue is the unawareness of the algorithms' presence. The Facebook News Feed is often presented as an example of algorithmic invisibility. Facebook's News Feed algorithm uses thousands of signals to display users the stories the system thinks are the most relevant (Facebook, 2016a). As a result, not all posts are displayed on the users' feed, a fact that is unbeknownst to many. In a study conducted with 40 Facebook users, Eslami et al. (2015) found that 62.5% of participants didn't knew that their Facebook News Feed was actively managed by an algorithm. Other studies indicate that this figure can be higher than 75% (Hamilton et al., 2014). Eslami et al. (2015) underline the potential social implications of the algorithm's invisible hand: "participants used News Feed to make inferences about their relationships, wrongly attributing the composition of their feeds to the habits or intent of their friends and family.".

In order to understand how bias and mistakes can make their way into algorithms, it is important to grasp how these constructs make decisions.

2.2.1 How Algorithms Decide

Diakopoulos (2014, 2016) describes the multiple steps involved in the decision-making process of algorithms.

Prioritization is employed to determine which criteria or metrics to emphasize in detriment of others. Since there are choices involved in which criteria to prioritize, such algorithmic decisions can introduce bias.

Classification is the process that categorizes an entity as pertaining to a given class based on the entity's features. There can be uncertainty in categorizing an entity one way or another, leading to classification errors and bias. Theses mistakes are known as false positives (an entity belongs to class A but is classified as B) and false negatives (an entity belongs to class B but is classified as A). Even though algorithms can be tuned to make fewer mistakes of a single type, this has implications, since it often results in more mistakes of the other type.

Association is the process of creating relationships between entities. Similarly to classification, association decisions can suffer from false positive and false negative errors. Another issue

is related to how people can interpret those associations, i.e., the association of an entity with a negatively connoted one can be perceived as a causal link rather than simple correlation.

Filtering involves the inclusion or exclusion of information based on various rules or criteria. Information can be filtered in or out according to previous prioritizing, classification, or association decisions. Filtering decisions can overemphasize or exclude certain information, and therefore potentially contribute to create filter bubbles, i.e., amplify biases and prevent exposure to diverse perspectives by showing information with which people already agree.

While the decision-making process of algorithms might be conceptually simple to fathom, the decisions, choices, and influences embedded into them by humans are often undisclosed, making them akin to "black boxes" (Diakopoulos, 2014). The rise of algorithm usage in news media, and the potential social implications it entails, have led to a call for algorithmic transparency and accountability.

2.2.2 Algorithmic Transparency

Research about how systems can be transparent about the use of algorithms is still at an embryonic phase. There are no agreed upon standards, and regulation is still lacking (Diakopoulos, 2016). Some attempts to fill this gap have been made. The Association for Computing Machinery US Public Policy Council (2017) recently published a press release listing a set of recommendations to increase algorithmic accountability and transparency. The document mentions seven principles:

Awareness All stakeholders of algorithmic systems should be aware of the potential biases embedded in the code and the possible negative outcomes to individuals and society.

Access and redress Algorithmic systems should provide mechanisms through which individuals and groups negatively impacted can question and rectify the outputs.

Accountability Institutions using algorithms in their decision making process should be held responsible for those decisions.

Explanation Explanations about the algorithms' procedures and the specific resulting decisions should be provided.

Data Provenance Information about how the training data was collected and the exploration of potential biases induced by the data-gathering process should be maintained. In order to protect privacy, intellectual property and preventing gaming the system, access to this information can be restricted to qualified and authorized individuals.

Auditability Information about the models, data, algorithms, and decisions, should be recorded in order to allow its auditing when harm is suspected.

Validation and testing Models should be rigorously validated, particularly to assess if they can be harmful, and the process documented and, ideally, made public.

Research with representatives from the news field has also resulted in some recommendations. After conducting a study with 50 participants from the news media industry and academia, Diakopoulos and Koliska (2016) identified four layers of algorithmic systems on which transparency factors might be disclosed: data, model, inference, and interface. A summary of the authors' findings is presented in Table 2.1. Diakopoulos and Koliska (2016) also mention two major challenges to algorithmic transparency in news media: i) the lack of financial incentives (e.g., the costs imparted on producing transparency information, loss of competitive advantages, legal issues, or opening the door to manipulation), and ii) the concern of negatively affecting the user experience by presenting too much information.

2.2.3 User Experience of Algorithmic Transparency

As mentioned previously, the topic of algorithmic transparency is still in its infancy, resulting in a deficit of normative approaches. However, some starting points have been provided in the scholarly literature.

Hamilton et al. (2014) proposed three approaches to investigate algorithmic awareness and its influence on user behaviour: "(1) surveying users to determine their awareness of processes at work in their everyday consumption; (2) exposing hidden algorithmic processes to users and then studying the effects of knowledge on use; and (3) working with users to try and deduce the algorithmic processes at hand, as well as the design rationales behind them".

In effect, some of these approaches, individually or combined, have been employed by some studies. Eslami et al. (2015) investigated the issues of algorithmic awareness by showing participants the difference between two versions of their News Feed, one curated by Facebook's algorithm and one without curation. The authors found that the study's participants unaware of the curation (62.5%) reacted negatively after discovering the algorithm's presence. However, after some time, knowledge about the algorithm increased satisfaction with the product. Based on these findings, the authors suggest that disclosing the algorithm's existence can give users a sense of agency and control, thus enhancing the experience.

Similarly, Kizilcec (2016) found that a user interface (UI) that promotes algorithmic transparency can elicit positive attitudes, namely increase the users' trust towards the system. Nevertheless, the author found that too much transparency can potentially negate the positive effects, which might indicate "a bell-shaped relation between transparency and trust".

Diakopoulos and Koliska (2016) research with representatives from the news media industry and academia also provides some insights on how algorithmic transparency can be conveyed through UIs. These include signalling algorithmically processed content by displaying an icon next to it; allowing the manipulation and tweaking of the algorithm's input parameters — as well as the ability to turn it on and off —, and visualizing the output; or providing interactive cues linking to textual descriptions of the algorithm's operation.

Layer	Factors

Data	• Information quality.
	  – Accuracy.
	  – Uncertainty (e.g. error margins).
	  – Timeliness.
	  – Completeness.
	• Sampling method.
	• Definitions of variables.
	• Provenance (e.g. sources, public or private).
	• Volume of training data used in machine learning.
	• Assumptions of data collection.
	• Inclusion of personally identifiable information.

Model	• Input variables and features. • Target variable(s) for optimization. • Feature weightings. • Name or type of model. • Software modeling tools used. • Source code or pseudo-code. • Ongoing human influence and updates. • Explicitly embedded rules (e.g. thresholds).
Inference	• Existence and types of inferences made. • Benchmarks for accuracy. • Error analysis (including e.g. remediation standards). • Confidence values or other uncertainty information.
Interface	• Algorithmic presence signal. • On/off. • Tweakability of inputs, weights.

Table 2.1: Summary of transparency factors across four layers of algorithmic systems (Diakopoulos and Koliska, 2016)

A study with thousands of participants from all over the world found that users were strongly concerned that algorithmic news applications might not expose them to important information and challenging perspectives, and also jeopardize their privacy (Newman et al., 2016). Algorithmic transparency can potentially assist in alleviating those worries.

Some of the insights mentioned above can be seen in action on the IEEE Spectrum Top Programming Languages Interactive Ranking web application[10] (Diakopoulos et al., 2014). As shown in Figure 2.2, the application's UI allows users to manipulate the algorithm that generates the ranking by changing the weightings of each data source, and then check the effect of these alterations. The system also permits a side-by-side comparison of two rankings, displaying a visualization of the ordering differences between them (Figure 2.3). A web page detailing the sources of data and the methods used is also available[11].

According to Diakopoulos (2016), about one in six of the 1285 tweets about the application indicated that users were reweighting the ranking in multiple ways, which might be indicative of a successful design.

[10] http://spectrum.ieee.org/static/interactive-the-top-programming-languages-2016
[11] http://spectrum.ieee.org/ns/IEEE_TPL_2016/methods.html

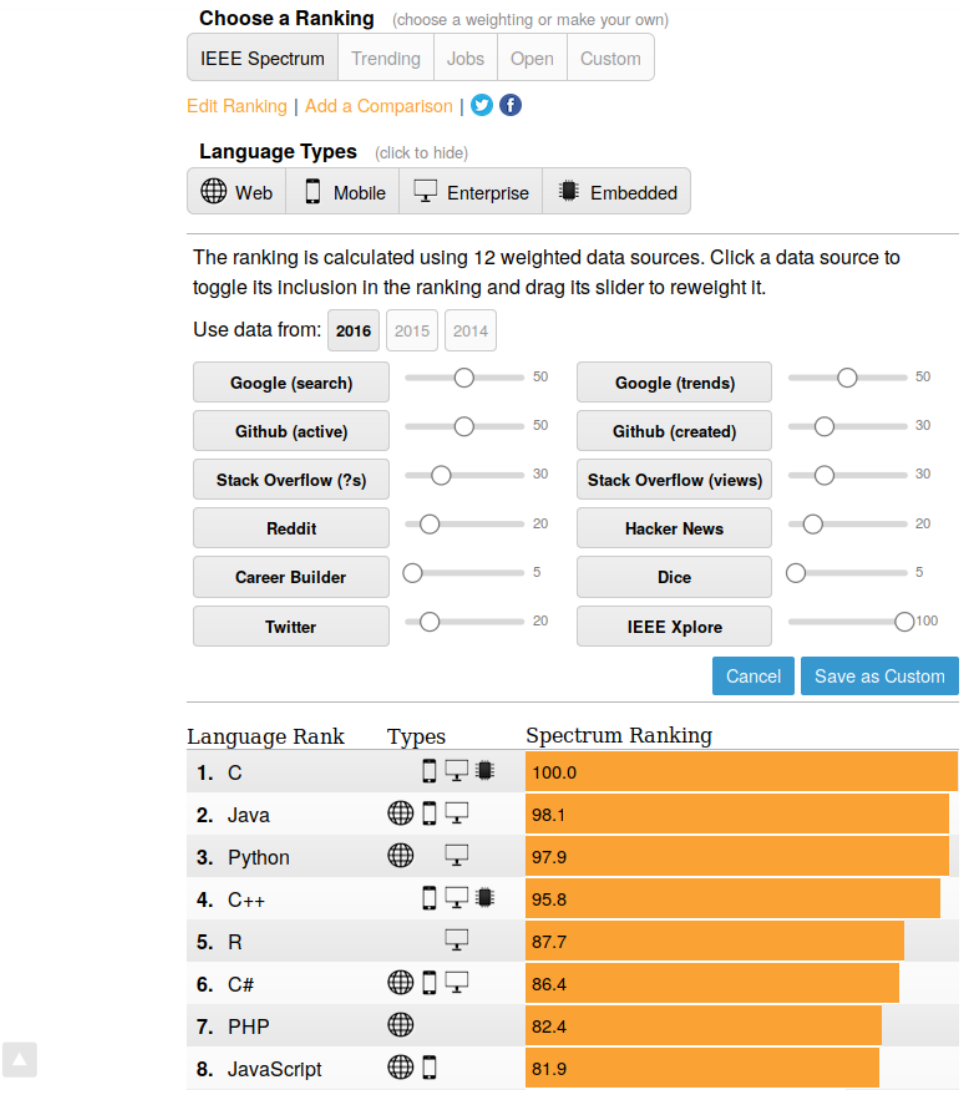


Figure 2.2: IEEE Spectrum Top Programming Languages Interactive Ranking: Weighting manipulation. Screen capture from http://spectrum.ieee.org/static/interactive-the-top-programming-languages-2016

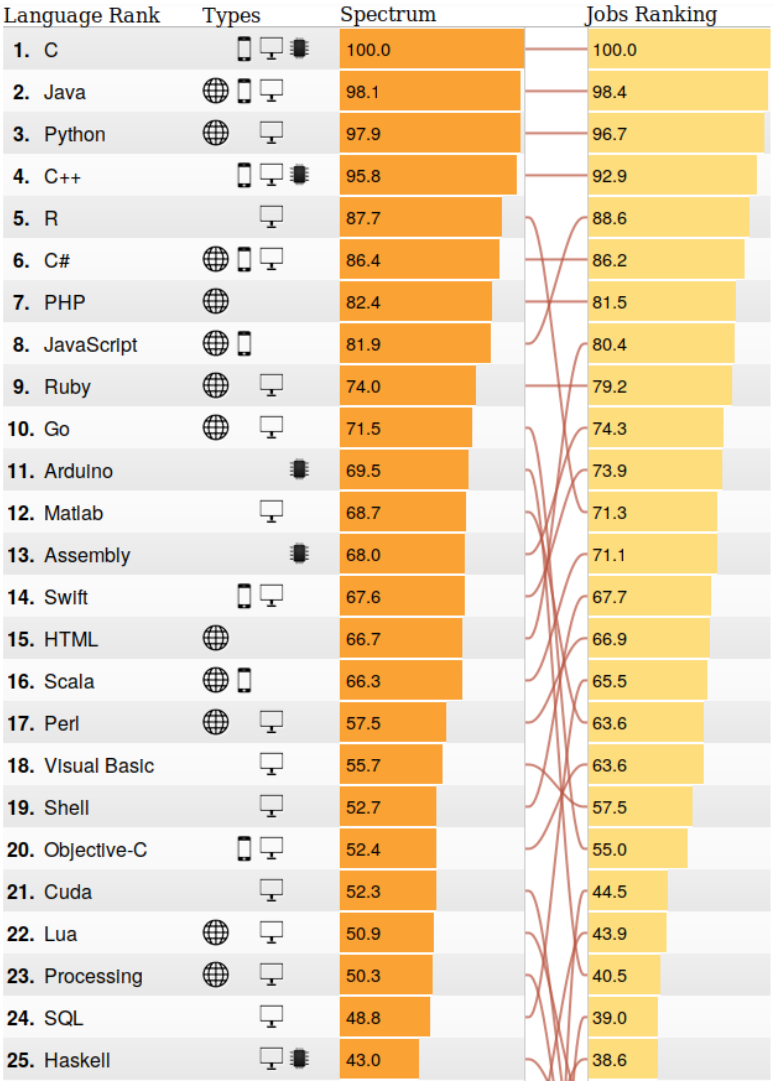


Figure 2.3: IEEE Spectrum Top Programming Languages Interactive Ranking: Ranking comparison. Screen capture from http://spectrum.ieee.org/static/interactive-the-top-programming-languages-2016

2.3 Summary

In this chapter, we defined the different types of news aggregators and identified some features that can improve the experience of end-users of these systems. Then, we presented the implications that the opaqueness and manipulation of algorithms can have in society. We reviewed the different steps that these computer programs use to make decisions, and presented the discussion on how they can be made more transparent. Finally, we surveyed some strategies to convey transparency and provide control to end-users of algorithmic systems. The literature reviewed in this chapter provided us with several guiding cues regarding the development of an algorithmic news aggregator. Our application won't recreate content and will focus on professionally produced news stories. According to the typologies surveyed it would therefore classified as a feed aggregator (Isbell, 2010) located in the lower left quadrant of the categorization presented by Coddington (2015) (Figure 2.1). It will also aim to provide a user friendly interface, high quality and reputable sources, and personalization features (Chowdhury and Landoni, 2006). Algorithmic transparency and control will be conveyed through the UI layer, using strategies such as signalling the algorithm's presence, allow users to change their news feed by manipulating the algorithm's parameters, and provide textual descriptions of how it operates (Diakopoulos and Koliska, 2016; Diakopoulos, 2016).

Chapter 3

Methods

In this chapter we start by describing the study design, the participants and their main demographic characteristics, the sampling process and its implications regarding representativeness. We then describe the data gathering instruments used and the procedure followed to conduct the experiment, and explain how the collected data was analysed. Finally, we characterize the sample in terms of news consumption habits.

3.1 Study Design

This study employed a non-experimental design, i.e., there was no independent variable manipulation nor a control group or multiple measures. A quantitative instrument – a structured questionnaire – was used to collect standardised data about the variables of interest.

3.2 Participants

Our sample was comprised of four hundred and thirty two (n = 432) participants (235 women, 194 men, and 3 who didn't disclosed the gender). They were divided in six age groups: 15-24 (n = 227), 25-34 (n = 141), 35-44 (n = 32), 45-54 (n = 18), 55-64 (n = 11), 65+ (n = 2). One participant didn't disclosed the age group. The sample mainly consisted of bachelor's, master's and doctorate students from multiple programmes, academic staff (primarily professors) of various ranks, and also some former students with a wide range of professions.

3.2.1 Sampling method

Participants were recruited via an email sent through the University of Porto's (U.Porto) information system email service. Due to administrative restrictions that limit whom students can contact via email, accessing the address list of the whole population – the U.Porto academic community – wasn't possible. This study's sample was therefore drawn after contacting the following groups: students and professors from the Communication Sciences Bachelor's and Master's programmes at the Faculty of Letters (FLUP), academic staff from the Faculty of Economics

(FEP), and to every student enrolled between 2009 and 2016 in FLUP, FEP, Faculty of Engineering (FEUP), Faculty of Fine Arts (FBAUP), and Faculty of Sciences (FCUP). No personally identifiable information was collected, so participants operated under complete anonymity.

Since no form of random selection could be employed, because of the aforementioned restrictions to the sampling process, this study used a purposeful non-probability sample. Thus, no assumptions can be made regarding the sample's representativeness.

3.3 Instruments

Data was gathered through a structured questionnaire written in Portuguese. The questionnaire was designed and hosted on the Google Forms[1] platform. For the purpose of this dissertation, the questionnaire was translated to English by the author. The English version of the full questionnaire can be found in Appendix A.1. The original version in Portuguese is available in Appendix A.2. Below we present the different variables investigated for each group and the count of questionnaire items.

- Demographics (age group, gender, occupation, scientific area of training, and academic degree) – 5 items
- News consumption habits (frequency of consumption, level of interest in news, type of media used to get news) – 3 items
- Level of importance of criteria regarding the news content – 17 items
- Level of importance of criteria regarding the diversity of perspectives in the news – 2 items
- Level of importance of criteria regarding the sources that produce the news – 5 items
- Level of importance of criteria regarding the information about the algorithm – 2 items
- Level of importance of criteria regarding the personalization of the news feed – 5 items

A Likert five-item scale, with the levels "Very Important', "Important", "Moderately important", "Little important", and "Not at all important", was employed for rating the level of importance of the criteria. An additional level, labelled "I don't know/Won't answer", was added to allow explicit non-responses. The questionnaire was piloted with four researchers. Based on their feedback, some redundancies in the descriptive text were eliminated, but no changes were made to the items, considered clear and adequate. No instrument was found in the literature susceptible of being adapted to the purposes of this study. The questionnaire employed was developed specifically for this study, and therefore not validated. However, the majority of the criteria were drawn from the literature. Table 3.1 lists the criteria by group and the corresponding references.

[1] https://forms.google.com

Table 3.1: List of criteria by group and corresponding references

Group	Criteria	References
News content	• Conflict • Good news • Shareability • Drama • Magnitude • Relevance • Surprise • Power elite • Audio-visual • Entertainment • Celebrity • Exclusivity • Bad news • Follow-up	Harcup and O'Neill (2016)
	• Proximity • Serendipity	Bozdag (2013)
Diversity	• Information that conforms to my beliefs • Information that presents competing points of view	Bozdag (2013)
Sources	• Perceived journalistic standards (adapted as journalistic prestige) • Size of news operation • Circulation statistics (adapted as reach)	Powers (2017)
	• Professionalism	Wendelin et al. (2017)
Personalization	• Explicit personalization • Implicit personalization based on my actions • Implicit personalization based on my friends' actions • Implicit community-based personalization	Bozdag (2013); Powers (2017)

3.4 Procedure

A link to the questionnaire was included in the recruitment email, which also mentioned the study's context, goal, and the approximate time needed to fill (based on the pilot test). The English translation of the email sent to the participants is in Appendix B.1, and the original Portuguese in Appendix B.2. Further instructions and information were included in the questionnaire's descriptive text. An effort was made to provide examples and clarify potentially confusing concepts, and avoid technical jargon. Responses to the questionnaire were collected during a period of 34 days (from April 27 to May 30, 2017) and automatically saved in an online spreadsheet, a feature provided by the platform used to create the questionnaire.

3.5 Data Analysis

Data analysis was performed using a R programming language[2] script developed by the author in the RStudio Integrated Development Environment[3]. Questionnaire data was published to an URL in CSV format and then imported to RStudio. In addition to some data cleaning operations, such as removing irrelevant columns and the four duplicate responses identified, a set of functions was implemented to convert the variables into factors. Factors are a data object in R which categorize the data as a finite number of numeric levels and therefore allow the statistical analysis of the two data types collected: nominal and ordinal variables.

A set of functions to plot the processed data was also developed. The mean was chosen as the measure of central tendency due to the considerable sample size. Pairwise group comparisons were performed using the Pearson's Chi-squared test for nominal variables, and the Mann-Whitney test for ordinal data. The choice of a non-parametric method was due to the fact that the data was found to not follow a normal distribution. An alpha of 0.05 was selected for rejecting the null hypothesis that there weren't significant differences between groups.

3.6 Sample Characterization

In this section we characterize the study's sample in terms of news consumption habits. They were assessed through three questionnaire items: frequency of news consumption, level of interest in news, and type of media used to consume news.

It was found that the majority of the participants gets news from television (71.2%), social networks (70.5%), and the media organizations' own websites and applications (69%), consumes news several times per day (57.3%), and is interested (45.9%) or very interested (26.3%) in news.

News consumption habits were also analysed in terms of age and gender groups. In order to explore the effect of age in more generational terms, the six original age groups were recoded into two aggregating groups: 15-44, comprised of every participant from the 15-24, 25-34, and

[2] https://www.r-project.org/
[3] https://www.rstudio.com/

35-44 groups, and 45-65+, with all participants from the 45-54, 55-64, and 65+ groups. Some variation was found between age and gender groups regarding the media type used to consume news, interest in news, and frequency of news consumption.

Participants from the 15-44 group mostly get their news from social networks (72.4%), television (70.2%), and the media's own websites or applications (69.3%). The 45-65+ group primarily get news from television (76.7%), the media's websites or applications (70%), and equally from print and radio (56.7%). This is shown in Figure 3.1. Regarding the interest in news, displayed in Figure 3.2, the majority of the respondents from the 15-44 group are interested (45.9%) or very interested (24.8%) in news, while most of the participants from the older group indicated that they are equally very interested or interested in news (46.7%). As for frequency of consumption, shown in Figure 3.3, the 15-44 and 45-65+ groups mainly consume news several times per day, albeit in different proportions – 56.6% and 66.7%, respectively.

Regarding the gender, women get most of their news from television (76.6%), social networks (74.5%), and the media's websites or applications (66%). Men primarily get news from the media's websites or applications (72.7%), social networks (65.5%), and television (63.4%), as shown in Figure 3.1. Women are mainly interested (48.1%) or moderately interested (24.7%) in news, while men indicated to be interested (43.3%) or very interested (32.5%) (Figure 3.2). The frequency of consumption between genders, displayed in Figure 3.3, shows that the majority of women and men consumes news several times per day – 55.3% and 59.8%, respectively.

In order to examine the significance of group differences beyond the tendencies presented above, inferential statistical tests were performed – Chi-squared tests for nominal variables (type of media), and Mann-Whitney tests for ordinal variables (interest in news and consumption frequency). Differences between men and women were found to be statistically significant regarding all measured dimensions – type of media used to get news, interest in news, and consumption frequency ($p < 0.05$). Gender differences across all levels were also inspected, via two-proportions z-tests. A two-tailed z-test indicated that the proportion of media type used by gender was significantly different for television, social networks, aggregation websites/applications and print media ($p < 0.05$). Results from one-tailed z-tests showed that, proportionally, at the 0.05 significance level, women get significantly more news from television and social networks than men. On the other hand, the proportion of men who consume news through print, blogs, and aggregation sites/applications was found to be significantly higher than women's at the 0.05 significance level. A two-tailed z-test showed that the proportion of interest in news by gender was significantly different at the "Very interested" level. The proportion of men who declared to be very interested in news was found to be, via a one tailed z-test, significantly higher than the women's. Regarding the frequency of consumption, a significant difference in proportions was found at the "several times a week" level, which a subsequent one-tailed z-test showed to be significantly higher in women.

Differences between age groups were found to be significant in terms of interest in news ($p < 0.05$), but not regarding the media type and consumption frequency. Interest in news by age group, in terms of proportions, was found to be significantly different at the "Moderately interested" and "Very interested" levels by a two-tailed z-test. A subsequent one-tailed z-test

indicated that the proportion of participants from the 15-44 age group moderately interested in news was significantly higher than the 45-65+ group. Conversely, a significantly greater proportion of participants from the 45-65+ group was found to be very interested in news.

In order to assess if our results were in line with existing works, we compared them with the findings from a report about the socio-demographic profile of Portuguese news consumers, conducted with 1049 participants and based on data from 2015 (Obercom, 2016). Generally, results from this study reflect the report's findings. Regarding gender differences, these include the fact that, when compared to men, women resort more to social networks and consume less news. In terms of age differences, the data from Obercom (2016) also mirrors our finding that older people are more interested in news than younger generations.

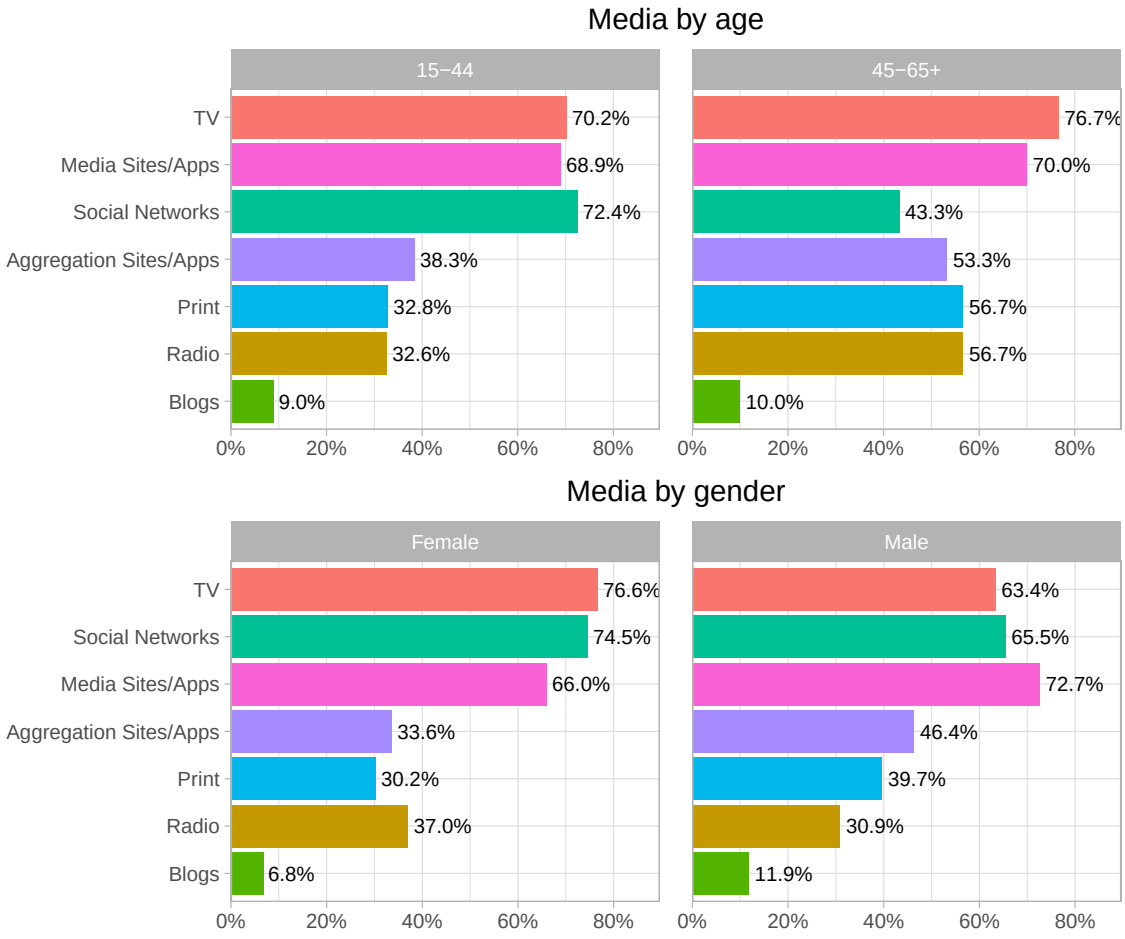


Figure 3.1: Media type used for consumption by age and gender

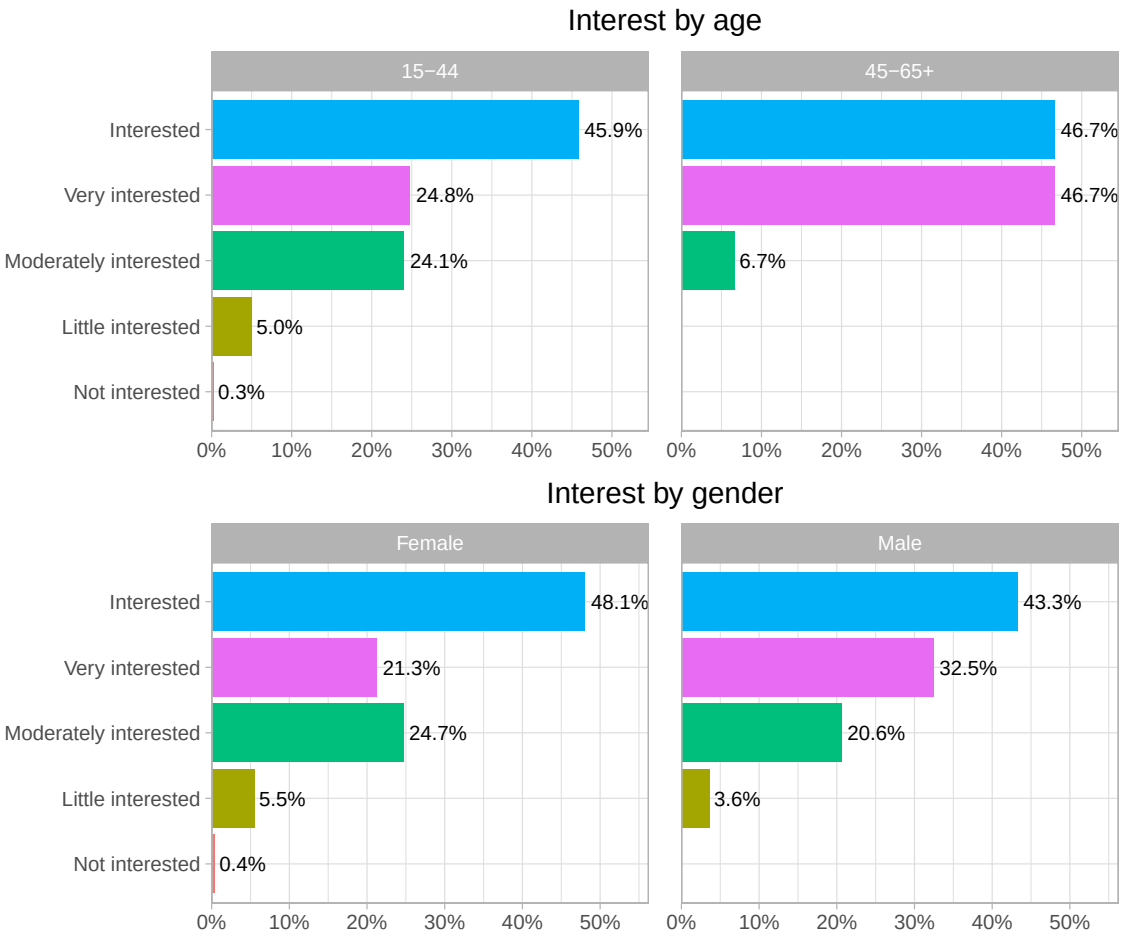


Figure 3.2: Interest in news by age and gender

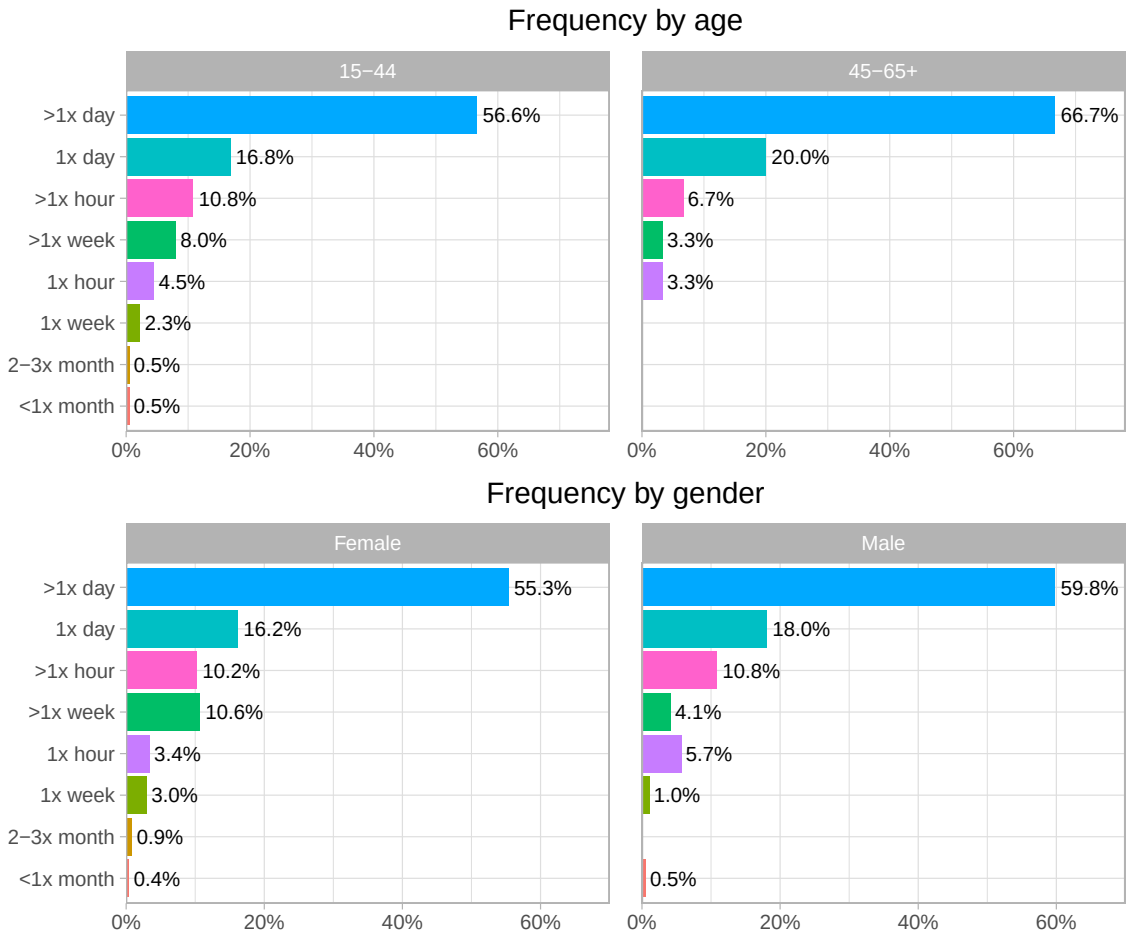


Figure 3.3: News consumption frequency by age and gender

3.7 Summary

This chapter presented the methodology employed in this non-experimental study. A sample of 432 participants from the University of Porto filled a questionnaire measuring demographics, news consumption habits, and the level of importance of a set of criteria, mostly drawn from the literature, related with news content, diversity, sources, algorithmic information, and personalization.

Data was analysed using a R script and the sample characterized in terms of news consumption habits: preferred media to get news, interest in news, and frequency of consumption. Participants indicated that they mainly get news several times per day from television, social networks and the publishers' own websites or applications, and are interested or very interested in news. Differences between age and gender groups were also investigated. They were found to be statistically significant ($p < 0.05$) in terms of the preferred media types, interest in news, and consumption frequency between men and women.

It was found that women get significantly more news from television and social networks than men, who display a significantly higher consumption of news through print, aggregation sites/applications and other media types. A significantly higher proportion of men declared to be very interested in news, and significantly more women indicated a consumption frequency level of several times a week. Regarding the two age groups examined – 15-44 and 45-66+ –, a significant difference was found regarding the interest in news. The proportion of participants from the 15-44 age group who were moderately interested in news was significantly higher than the 45-65+ group. Participants from the 45-65+ group, in turn, were found to be very interested in news in a significantly greater proportion than ones from the younger group. These results are mostly in line with a previous work about the socio-demographic profile of Portuguese news consumers.

Chapter 4

Results

This chapter presents the data analysis results of the questionnaire items measuring the level of importance of the criteria related to news content, diversity, sources, algorithmic information, and personalization. Values corresponding to the non-response level ("I don't know/Won't answer") were discarded from the analysis. Therefore, all the charts in this chapter present the number of valid observations below each item label in the format ($n = \#observations$). Additionally, each chart presents two horizontal axes which should be read independently. The bottom one displays the percentage of responses for each importance level in decreasing order of importance. The top axis is meant to assist the interpretation of each item's mean value. The numeric values displayed correspond to the importance level in ascending order according to a scale where a score value of 1 corresponds to the attribute "Not important", 2 to "Little important", 3 to "Moderately important", 4 to "Important", and 5 to "Very important". Each chart is accompanied by a table listing the questionnaire items for each criteria group and the corresponding label as it was displayed to the respondents, including the examples and clarifications provided.

4.1 News Content Criteria

The importance of the criteria related to news content was measured by 17 items. These items and the corresponding labels are presented in Table 4.1. On average, participants rated "Timeliness" (Q10), with a mean score value of 4.16 (SD=0.79), and "Good news" (Q14), averaging 4.13 (SD=0.88), as the most important news content criteria. More than 75% of the respondents evaluated these two criteria as very important or important, placing them above the 4th importance level, as shown in Figure 4.1. The conversion of the mean score values to the corresponding scale attribute indicates an average rating of "Important".

The criteria "Relevance" (Q18) (M=3.88, SD=0.82), "Conflict" (Q12) (M=3.78, SD=0.9), "Magnitude" (Q17) (M=3.76, SD=0.88), "Serendipity" (Q13) (M=3.7, SD=0.83), "Audio-visual" (Q21) (M=3.65, SD=0.96), and "Proximity" (Q11) (M=3.6, SD=0.94), were found to be, on average, closer to the 4th than to the 3rd importance level. The corresponding scale attribute of

these score values is also “Important”, albeit lower than the previous group in the hierarchy of importance.

Next are the “Follow-up” (Q26) (M=3.37, SD=0.89), “Power elite” (Q20) (M=3.31, SD=0.97), and “Surprise” (Q19) (M=3.15, SD=0.92) criteria. Their mean score values place them above of the 3rd importance level, whose corresponding attribute is “Moderately important”. The criteria “Bad news” (Q25) (M=2.96, SD=0.94), “Exclusivity” (Q24) (M=2.94, SD=1.09), “Entertainment” (Q22) (M=2.84, SD=1.01), and “Drama” (Q16) (M=2.69, SD=1.03), were also close to the “Moderately important” level. Finally, near to the 2nd importance level, whose corresponding attribute is “Little important”, are the criteria “Sharing potential” (Q15) (M=2.15, SD=1.04), and “Celebrities” (Q23), which, with a mean score value of 1.7 (SD=0.9), was the least important criterion.

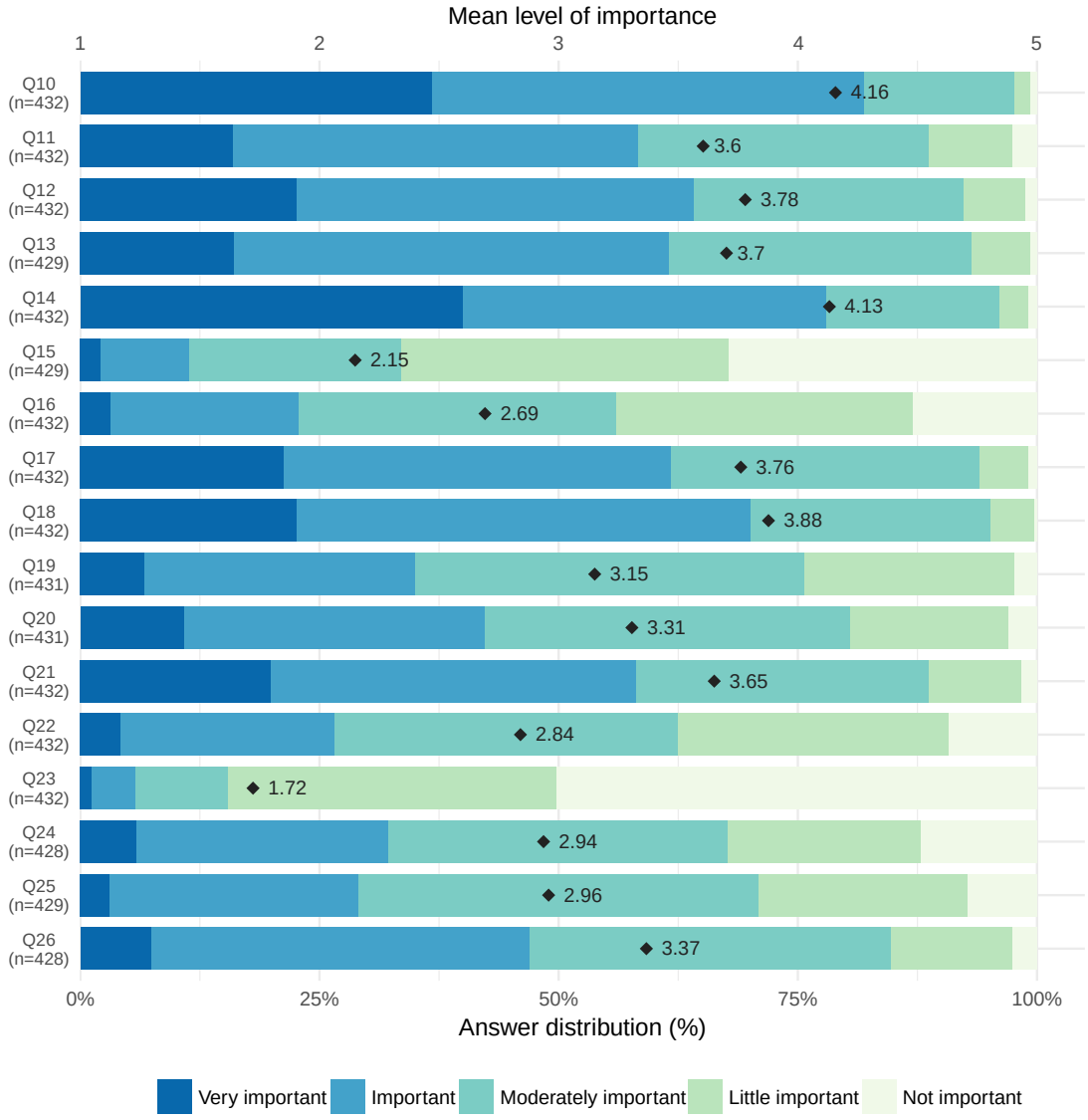


Figure 4.1: News content criteria

Table 4.1: News content criteria item labels

ID	Label
Q10	Timeliness (news published on top of the moment)
Q11	Proximity (news about events near where I am)
Q12	Conflict (news about controversies, strikes, wars)
Q13	Serendipity (discovering interesting information unexpectedly)
Q14	Good news (news with positive tones - cures, scientific advances, recoveries, victories and celebrations)
Q15	Sharing potential (news potentially able of generating shares and comments on Facebook, Twitter, and other social networks)
Q16	Drama (news about a developing drama such as escapes, accidents, searches, rescues, court cases)
Q17	Magnitude (important news due to the large number of people involved or potential impact)
Q18	Relevance (news about groups or nations understood as influential or historically and culturally similar to the audience)
Q19	Surprise (news with an element of surprise, contrast, out of the ordinary)
Q20	Power elite (news about powerful entities such as individuals, organizations, institutions or companies)
Q21	Audio-visual (accompanying news from photos, videos, audio and interesting graphics)
Q22	Entertainment (light news - sex, sport, show business, animals, humorous treatment)
Q23	Celebrities (news about famous people)
Q24	Exclusivity (original news and first published in a source)
Q25	Bad news (news with negative tones - deaths, injuries, defeats)
Q26	Follow-up (news about subjects already present in the news)

4.2 Diversity Criteria

Diversity criteria importance was evaluated through the two items presented in Table 4.2. As shown in Figure 4.2, the importance of being exposed to news that present different perspectives and points of view from their own (Q29) was evaluated by the participants as the most important criterion of the two, with a mean score value of 3.79 (SD=0.79), and a corresponding scale attribute of "Important". The importance of exposure to news in line with their perspectives, was rated on average as being "Moderately important", with an average score value of 3.36 (SD=0.94)

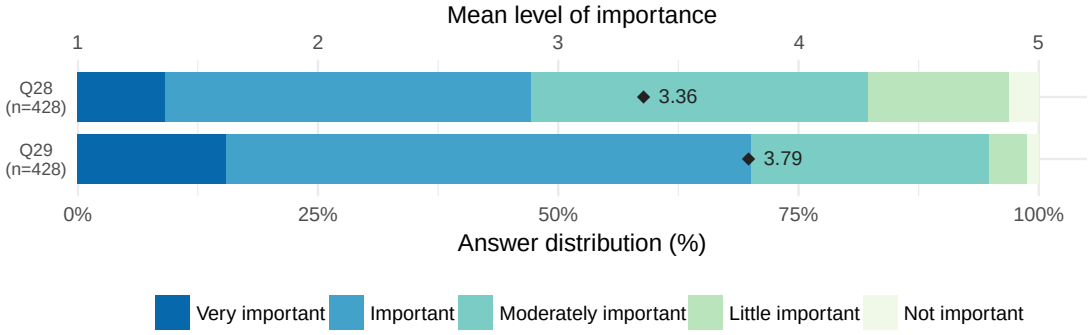


Figure 4.2: Diversity criteria

Table 4.2: Diversity criteria item labels

ID	Label
Q28	News that are in line with my perspectives and views
Q29	News that present different perspectives and points of view from mine

4.3 Sources' Criteria

Responses to the sources' criteria items, presented with their corresponding labels in Table 4.3, show that "Professionalism" (Q32) (M=4.34, SD=0.87) and "Journalistic prestige of the source" (Q30) (M=4.32, SD=0.79) were on average the two most important criteria, with an interpreted attribute of "Important". This can be seen in Figure 4.3. The criteria "Reach of the source" (Q33) (M=3.33, SD=1) and "Size of the source's operation" (Q31) (M=2.89, SD=1.04) correspond to the "Moderately important" attribute.

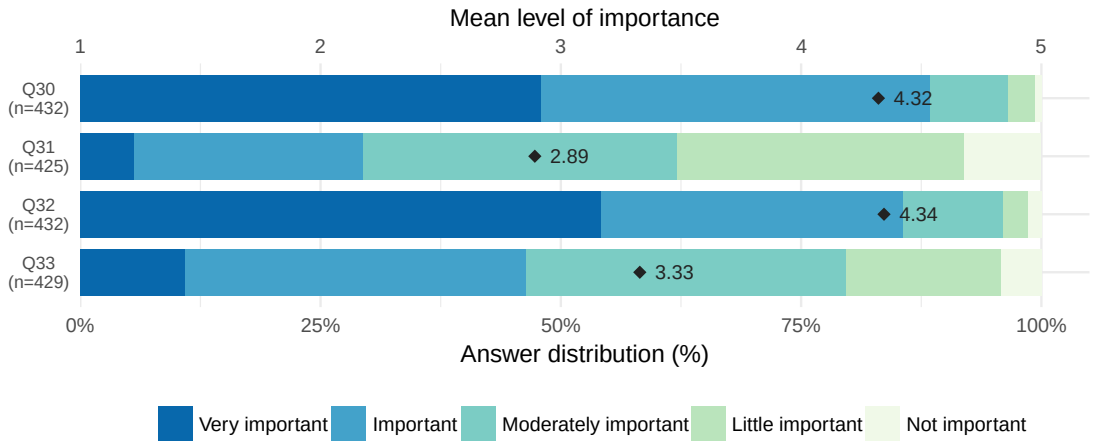


Figure 4.3: Sources' criteria

Table 4.3: Sources' criteria item labels

ID	Label
Q30	Journalistic prestige of the source
Q31	Size of the source's operation (the number of people working there)
Q32	Professionalism of the source (if the news are written by professional journalists)
Q33	Reach of the source (if it reaches many people)

4.4 Algorithmic Information Criteria

Regarding the two items for algorithmic information criteria, shown in Table 4.4, they were evaluated, on average, as "Moderately important". As displayed in Figure 4.4, the importance of having an indication that an algorithm was responsible for selecting the news (Q35) (M=3.3, SD=1.17) and information about the various steps used by the algorithm to achieve the selection (Q36) (M=3.32, SD=1.1) was rated similarly.

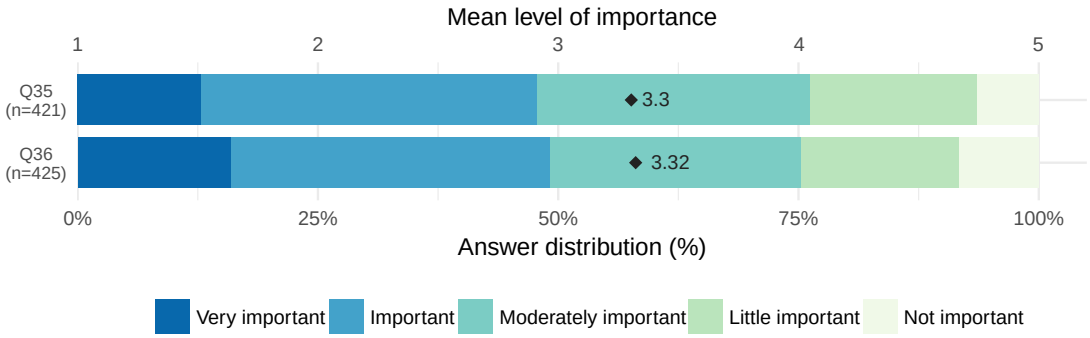


Figure 4.4: Algorithmic information criteria

Table 4.4: Algorithmic information criteria item labels

ID	Label
Q35	Indication that news selection is done by a computer program
Q36	Information about the various steps used by the computer program to select the news presented

4.5 Personalization Criteria

Five items were employed to assess the importance of the personalization criteria. They can be found in Table 4.5. As shown in Figure 4.5, the ability to configure a news algorithm according to the user's criteria – "Explicit personalization" (Q38) –, was rated on average as having the highest importance, with a mean score of 3.69 (SD=1.1) and a corresponding attribute of "Important". The criteria implicit personalization based on the user's actions (Q39) (M=3.01, SD=1.19) and no personalization (Q37) (M=2.92, SD=1.26) were rated closely, with a scale attribute of "Moderately important". Finally, the criteria related with implicit personalization based on the community's actions (Q41) (M=2.31, SD=1.1) and on the actions of friends (Q40) (M=2.15, SD=1.04) were judged on average as being "Little important".

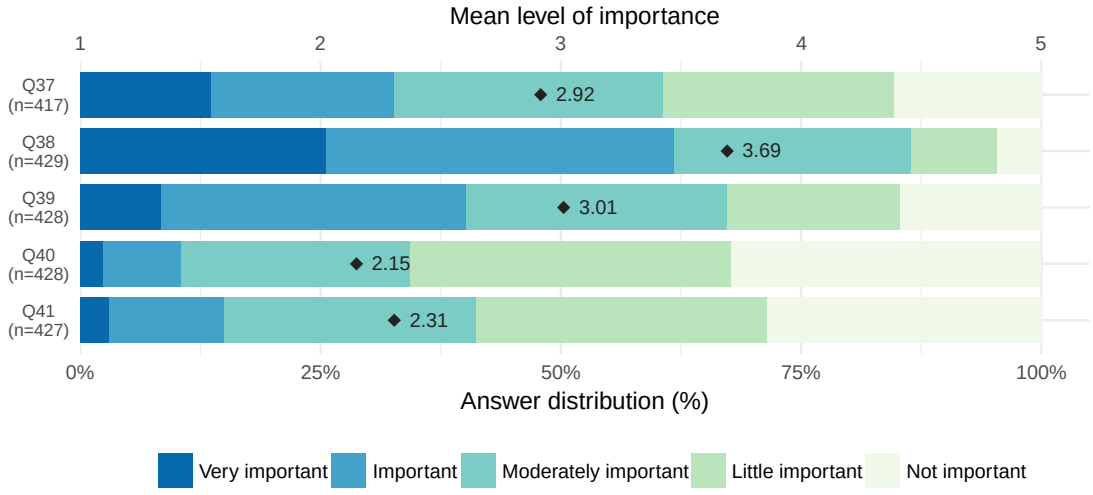


Figure 4.5: Personalization criteria

Table 4.5: Personalization criteria item labels

ID	Label
Q37	Without personalization (I do not want the program to create a news list adapted to my profile)
Q38	Explicit personalization (I want to be able to configure the program that generates the news list according to my criteria)
Q39	Implicit personalization based on my actions (I want the program to display news based on my action history - searches, clicks, comments, 'likes', etc.)
Q40	Implicit personalization based on my friends actions (I want the program to display news based on my friends' action history - searches, clicks, comments, likes, etc.)
Q41	Implicit community-based personalization (I want the program to display news based on the action history of most users - searches, clicks, comments, 'likes', etc.)

4.6 Summary and Discussion

In this chapter, we presented the results obtained via the analysis of the participants' responses to the different criteria assessed in the questionnaire. Regarding the news content, the most important criteria, rated on average as "Important" and in decreasing order of mean score, were "Timeliness", "Good news", "Relevance", "Conflict", "Magnitude", "Serendipity", "Audio-visual", and "Proximity". As mentioned in Section 3.3, most of the these criteria were drawn from Harcup and O'Neill (2016). The authors examined 711 lead news stories published in 10 United Kingdom's (UK) newspapers in 2014 to investigate the frequency with which the criteria appeared on them. Additionally, they also examined the 25 most frequently shared news stories by UK users on two social networks – Facebook and Twitter – in 2014 to investigate if there were differences between the news values selected by journalists and the audience.

In order to be able to make a direct comparison with the findings from Harcup and O'Neill (2016), we recreated our ranking of importance as if it only contemplated the same criteria identified by the authors. For instance, as "Timeliness" was not one of the criteria identified by the authors, we went down our hierarchy of importance to find the first one that was, in this case "Good news", and placed it in the position of the former, and so on. It should be noted that they identified a total of 10 criteria, but one of them – "Newspaper agenda" – wasn't used in our questionnaire. Table 4.6 presents a comparison of the hierarchy of importance attributed by this study's participants to each criteria with the frequency of appearance found by Harcup and O'Neill (2016) in news stories published by UK newspapers and shared by UK users on social media in 2014.

Table 4.6: Importance level of the news content criteria evaluated in this study compared with their frequency of appearance in newspaper stories and news shared on social media found by Harcup and O'Neill (2016)

Criteria	Importance level	Frequency (Harcup and O'Neill, 2016)	
		Newspaper stories	News shared on social media
Good news	1st	9th	=7th
Relevance	2nd	6th	5th
Magnitude	3rd	7th	6th
Follow-up	4th	4nd	4th
Power elite	5th	5th	9th
Surprise	6th	2rd	2nd
Bad news	7th	1st	3rd
Entertainment	8th	3rd	1st
Celebrities	9th	8th	=7th
Newspaper agenda	—	10th	10th

As it can be seen, there are marked discrepancies. For instance, while "Good news" was the most important criterion in our (reformulated) ranking, it was the least frequent in the UK's newspaper stories, and the second least frequent in the news shared by UK users on social networks. The second and third most important criteria according to our hierarchy, "Relevance" and "Magnitude", respectively, were also relatively infrequent both in newspaper stories (6^{th} and 7^{th}) and news shared on social media (5^{th} and 6^{th}). Another glaring case is "Bad news", the third least important in our findings but the most frequently found on newspapers and third on social media. "Entertainment" should also be mentioned, as it was placed eighth in our ranking but was the third most frequent in newspaper stories and the first in news shared on social media.

The first thing that should be mentioned when discussing these findings is that they result from the comparison of distinct measures, i.e., the participants' subjective assessment of criteria importance versus the frequency with which Harcup and O'Neill (2016) identified them in news stories in UK newspapers and shares on social media. Even though this study's participants were completely anonymous, we cannot forgo the possibility of some kind of self-report bias. In other words, the participants' subjective evaluations of importance might not be reflected in their actual news consumption behaviour. For instance, entertainment and celebrity stories were among the least preferred themes in a survey of 1049 Portuguese news consumers Obercom (2016), yet tabloid journalism displays strong audiences both off and online. Another possible explanation is related with the sample itself, which, being comprised of people with higher education levels, might not be representative of the general Portuguese population's media preferences. An additional hypothesis is that Portuguese and UK audiences have very distinct tastes regarding their media diet. Only further research, including conducting a similar analysis to the one performed by Harcup and O'Neill (2016) but with Portuguese news stories, would enable examining the causes behind these disparities. While investigating these – undoubtedly interesting – findings isn't part of this study's goals, it worth noting they might provide an interesting avenue of research for future works.

In terms of diversity, the importance of news presenting perspectives diverse from the participants' was typically evaluated as "Important", a higher level than the one attributed to the importance of exposure to news in line with their views, rated on average as "Moderately important". Professionalism and journalistic prestige, whose average scores correspond to the "Important" level, were found to be the most important criteria regarding the content producing sources. There are similar findings in previous works. For instance, news produced by professional media outlets have been found to be the most read, shared, and discussed stories on social networks (Newman et al., 2015).

Indication that the news selection was done by an algorithm and information about the various steps employed to achieve the selection, the two criteria related with algorithmic information, were typically found to be "Moderately important". We consider this to be an interesting finding, particularly the fact that they weren't more highly rated. There are indications that unaware users can react negatively when they find their news feed is managed by an algorithm, while, on the other hand, disclosing its presence can improve the experience (Eslami et al., 2015; Kizilcec,

2016). Concerns about the potential negative impact of algorithms have also been expressed by respondents from all over the world in a large scale study (Newman et al., 2016).

Finally, participants indicated that explicit personalization – the ability to configure the algorithms' output according to their own choices – was the most important personalization criteria, rated on average as "Important". So, participants value personalization, as found by Chowdhury and Landoni (2006), but they want to explicitly control it. A similar discovery is mentioned by Newman et al. (2016). After conducting focus groups in USA, UK, Germany and Spain, the authors state that "[...] it was clear that many active internet users now see themselves as editors – balancing and comparing multiple sources, multiple editorial judgements, and even multiple algorithms.".

Chapter 5

Implementation

This chapter discusses the implementation of the prototype developed in the context of this work. We present the news collection platform in which the prototype was integrated, and the two new components developed on top of the existing infrastructure. We then talk about how the findings from both the literature and the analysis of the participants' responses were integrated in the developed front-end application in order to achieve the proposed goals of alleviating information overload, providing user control, and present algorithmic transparency information.

5.1 System Description and Architecture

In order to have access to a vast archive of news articles on which to perform algorithmic aggregation, the prototype was integrated with the infrastructure of the MediaViz system (Devezas et al., 2015). MediaViz is a web platform that continuously collects and stores the articles published through the web syndication feeds of dozens of online sources. At the time of writing, the platform provided access to more than 4.3 million articles, collected since December 2014 from all major Portuguese media organizations, some international outlets and a few blogs.

MediaViz's data collection and access is performed by three conceptually distinct components. A Ruby on Rails[1] application runs a scheduler that regularly checks the feeds for new articles, saves them, and provides an administrative UI for managing the feeds to collect. A PostgreSQL[2] database stores all collected content and serves as the "single source of truth". The third component is an Apache Solr[3] index that is regularly synced with the database through a request sent by the Rails application.

For this work, two new components were developed and integrated on top of the existing infrastructure. The first was a Sinatra[4] web application that serves as a façade to the Solr index. It converts requests from a client application to the Solr query syntax, queries the index, and converts the payload to JavaScript Object Notation (JSON) before returning the article groups ("clusters").

[1] http://rubyonrails.org/
[2] https://www.postgresql.org/
[3] https://lucene.apache.org/solr/
[4] http://www.sinatrarb.com/

The clustering is performed by the open-source Lingo algorithm (Osiński et al., 2004), which was integrated with Solr and made accessible via an endpoint. The second component was a client front-end application, which provides the UI with the system.

Figure 5.1 shows the full system architecture after the integration of the developed components (marked as 1 and 2).

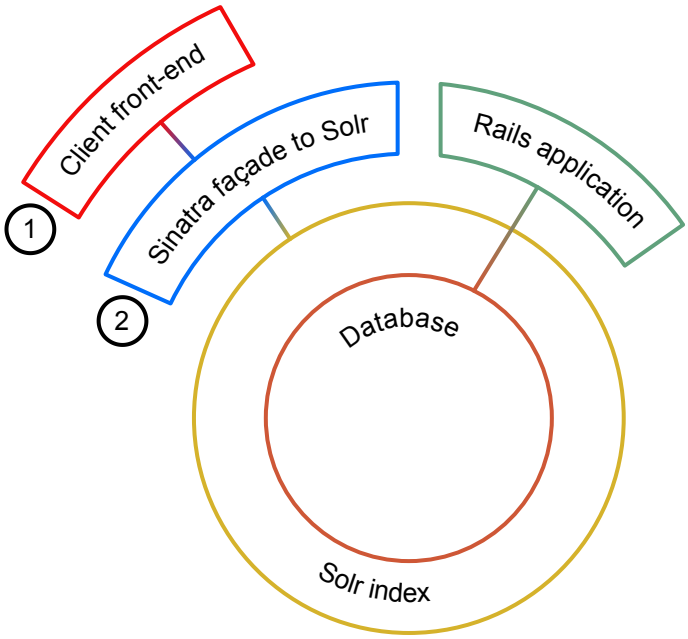


Figure 5.1: System architecture after the integration of the developed components

5.2 Front-end Application

End-users interact with the system through the front-end application, which was developed using web technologies – HTML, CSS, and the JavaScript framework Vue.js[5]. A live version is available in `http://clusters.tiagodevezas.pt`. As mentioned by Diakopoulos and Koliska (2016), the UI is one of the layers of algorithmic systems through which transparency can be conveyed. This can be achieved by factors such as signalling the algorithm's presence, provide links to textual descriptions about how it operates, and allow tweaking its parameters. These factors, paired with the requirements expressed by the participants, were taken into account during the development phase. Therefore, the developed application aims to achieve three main goals: i) combat information overload by algorithmically condensing thousands of articles into a few groups of related items about a given day's most relevant topics; ii) provide user control via explicit personalization; and iii) promote information transparency. We discuss below how each of these goals were translated to the front-end application.

[5] `https://vuejs.org/`

5.2.1 Condensing the News

When a user first enters the application, a news feed comprising a list of articles clustered by common topics is displayed, as shown in Figure 5.2. The article clusters are computed by the Lingo algorithm according to the default parameters and are comprised by a main article, whose title is shown with a larger font size, and, below, up to four related items covering the same topic. By default, the clusters displayed are respective to the current day, but users can click on the date to open a datepicker widget and select a different day — they can go back in time down to December 4, 2014, when the MediaViz system first started collecting news.

Figure 5.2: Front-end application home page

The start page thus provide an overview of the selected day's most relevant topics in a format that is easily digestible. An initial set of thousands of news articles is algorithmically processed and condensed into a few dozens article clusters. It should be mentioned that relevance in this context is an assumption, i.e., topics discussed in multiple articles are deemed to be more relevant than the ones mentioned in a single article. The clusters are scored by the algorithm according to the number of articles they contain, so the more articles a cluster contains, the higher the score. Due to this assumption, cluster score is the default ordering parameter. However, and based on the results from the questionnaire, where "Timeliness" was one of the two most important criteria regarding news content, the UI provides a drop-down button that allows ordering the clusters according to the publication time.

Another feature is the topic filter, placed on the left sidebar of the page. Users can click on a topic to only see the corresponding cluster, and click on it again to go back to the complete list. In addition to the compact presentation of the day's most relevant topics, this feature can help users to quickly focus on the topics that interest them without having to scroll the page. Finally, the system employs randomization to ensure that users are exposed to articles from different sources of information. When a cluster includes multiple articles, those articles are selected randomly. This means that two different users (or the same user, by performing a page refresh) can be presented distinct articles about the same topic. Figure 5.3 illustrates this feature: after a page refresh, the articles displayed in the cluster change, but the topic is the same. This can potentially expose users to multiple perspectives about a topic. The implementation of this feature is related with the results from the questionnaire, where participants indicated that being presented diverse points of views was the most important criterion regarding diversity.

Figure 5.3: Random selection of cluster articles

5.2.2 Controlling the Algorithm

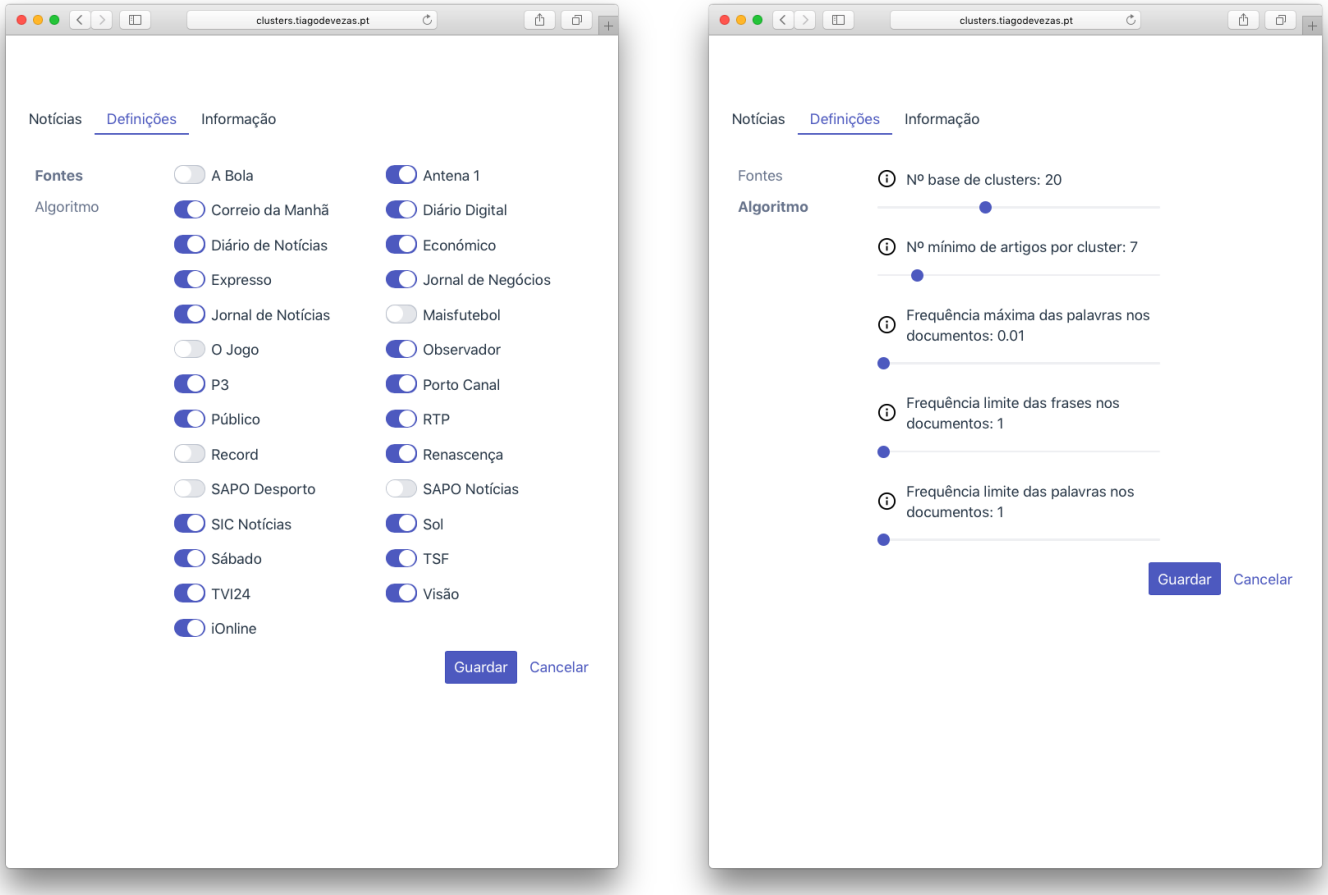
Participants indicated that explicit personalization, i.e., the ability to configure the algorithm according their own choices, was the most important criterion in terms of the desired degree of control over the system. The front-end application provides two ways to configure the algorithm and therefore manipulate how the presented list of news clusters is generated (Figure 5.4). The first way is the ability to toggle the sources from which the algorithm will derive the article clusters. This feature allows users to explicitly expand or contract the news feed according to their own interests and assessment of criteria such as the ones measured in the sources' segment of the questionnaire. As shown in Figure 5.4a, the source list comprises the major Portuguese professional media. Professionalism was evaluated by the participants as one of the two most important criterion regarding the sources, and it's also a highly valued feature of news aggregators (Chowdhury and Landoni, 2006).

Another way to control the system is to directly modify the parameters used by the algorithm, as displayed in Figure 5.4b. Users can get additional information about what each parameter does by hovering with the mouse cursor over the information icons. These settings allow changing the number of clusters displayed, the minimum number of articles per cluster, and also the frequency below which phrases and words are discarded from the algorithm's processing. While this represents just a subset of all the possible configurable parameters of the algorithm used, implementing more settings will be trivial, since all the groundwork has been laid. All the configurations set by the users are stored in the browser using the Local Storage web technology, so they are maintained between visits.

5.2.3 Providing Algorithmic Transparency

As discussed in Chapter 4, participants attributed moderate importance to the presentation of information about the presence and steps used by the algorithm to generate the news feed. Maybe users don't see algorithmic transparency as beneficial to user experience, a concern that has been expressed by news media representatives (Diakopoulos and Koliska, 2016). However, there are indications that it can give users a sense of agency and control (Eslami et al., 2015), and increase trust towards the system (Kizilcec, 2016), therefore enhancing the experience. On the other hand, an excess of transparency information can be detrimental (Kizilcec, 2016). In order to convey algorithm transparency in the developed application, three strategies mentioned by Diakopoulos and Koliska (2016) were employed: signalling that the feed is algorithmically processed, allow tweaking the algorithm's parameters (already discussed in the previous section), and linking to a textual description of the algorithm's operation. The indication that the content has been processed by an algorithm is displayed directly below the application's main header. This is shown in Figure 5.5.

The word "algorithm" is a link, so it serves a double purpose: in addition to indicate the algorithm's presence, when clicked, it sends users to the information page, displayed in Figure 5.6. This page provides a description of the steps followed by the algorithm to achieve the end result,



(a) Sources configuration settings

(b) Algorithm configuration settings

Figure 5.4: Application configuration settings

namely the preprocessing, phrase extraction, cluster content discovery, and final cluster formation phases. This page aims to provide sufficient detail to form a mental model of the algorithm's inner workings, without overloading the user with too much information and technical details. In order to contemplate cases in which a user may desire to deeply inspect and understand the system, some references from the algorithm's developers are provided.

Figure 5.5: Indication that the content has been algorithmically processed

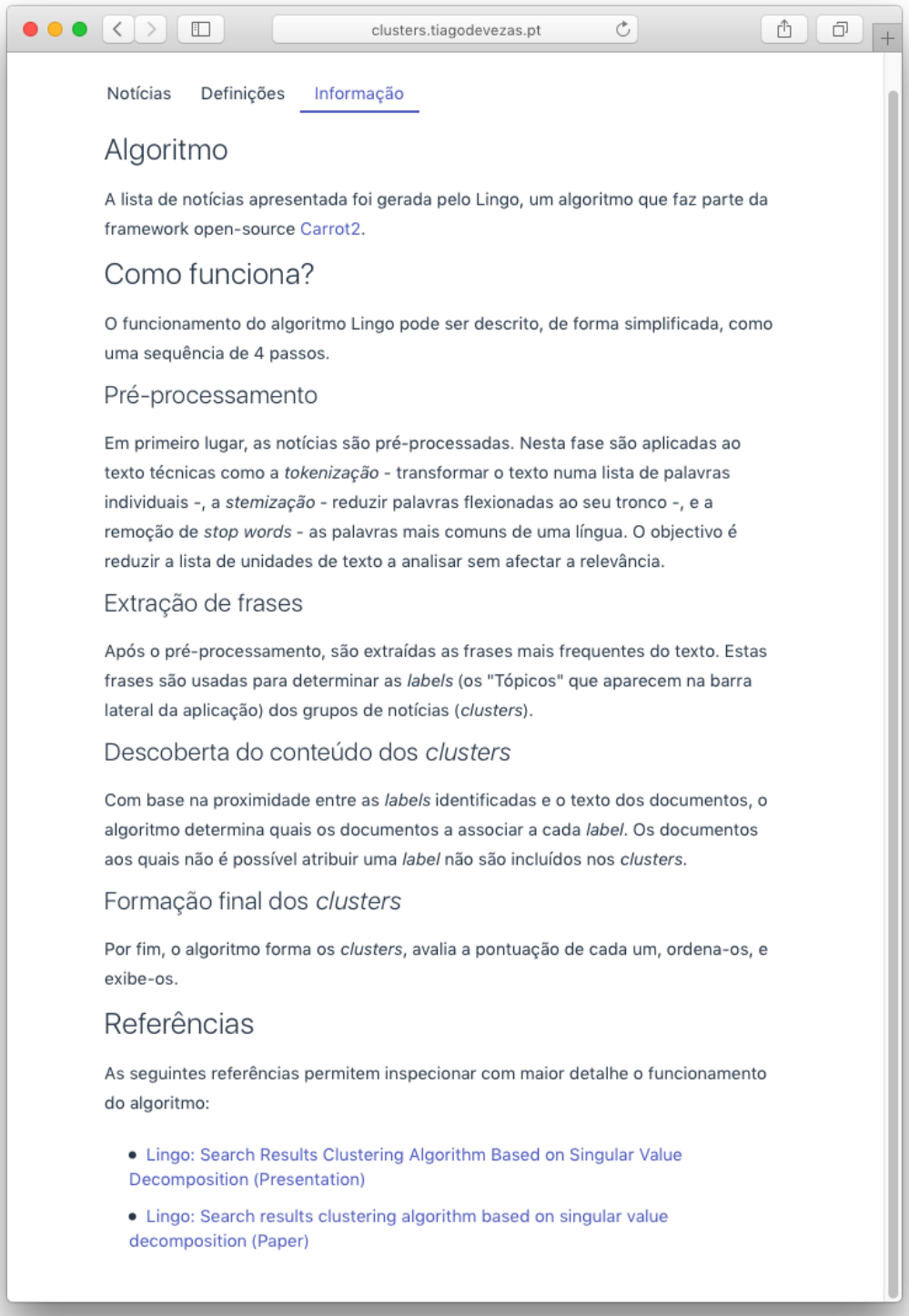


Figure 5.6: Information about the algorithm's operation

5.3 Summary

In this chapter, we described the implementation of a functional prototype of an algorithmic news aggregator concerned with user control and transparency. The application takes advantage of an existing project's infrastructure that provides access to a continuously updated archive of news articles. Two new components were developed and integrated on top of the previous architecture.

The implemented front-end application helps tackle information overload by condensing thousands of articles into a limited number of groups about the day's most relevant topics. It also aims to enhance diversity by randomly selecting the articles shown in each cluster. The system provides user control by allowing the explicit configuration of the sources and algorithm parameters used to generate the news feed. Transparency information about the algorithm's presence and operation is presented to the users, as well as in-depth references to allow deeper inspection.

Chapter 6

Conclusions

This study assessed the level of importance of multiple criteria regarding news content, diversity of perspectives, sources, algorithmic information, and personalization in the context of an algorithmic news application. Data was gathered using a questionnaire that was filled by four hundred and thirty two participants from the University of Porto community. Some of these findings, in conjunction with insights from the literature, were successfully applied to a functional prototype of an algorithmic news aggregation application. Personalization and curation algorithms help tackle information overload (Rader, 2017). But they can contain biases (Friedman and Nissenbaum, 1996), potentially lock users inside their ideological and cultural bubbles and help spread fake news, among other pernicious effects discussed in the literature. Despite their increasingly growing power over society, they often operate outside public scrutiny (Diakopoulos, 2014; Eslami et al., 2015; Hamilton et al., 2014). This has led to led to a call for greater transparency and accountability (Diakopoulos, 2014).

We found that the participants were aware of some of these issues. They indicated that the news content criteria they would most like to manipulate in such a system were "Timeliness", "Good news", "Relevance", "Conflict", "Magnitude", "Serendipity", "Audio-visual", and "Proximity". Some of the news content criteria were drawn from a study which also assessed how often they were used in UK newspapers stories and in the most shared news on social networks by UK users (Harcup and O'Neill, 2016). We found that the preferences expressed by this study's participants were very distinct from the ones identified by the authors in actual news stories. Regarding the diversity of perspectives, participants attributed higher importance to news that present points of view diverse from their own than to content aligned with their views. In other words, participants expressed that they not wish to be trapped inside filter bubbles. Professionalism and journalistic prestige were rated as the most important factors concerning the sources. This mirrors previous findings about the importance of professionally produced news content (Chowdhury and Landoni, 2006; Newman et al., 2015). Indication about the algorithm's presence and information about its operation was evaluated as being of moderate importance. The fact that these criteria weren't more highly rated was somewhat surprising, particularly due to the association of algorithmic intervention and the creation of filter bubbles, something that, as mentioned above,

participants indicated they would like to avoid. It's also at odds with the indications that unaware users can react negatively when they find their news feed is managed by an algorithm, while, on the other hand, disclosing its presence can improve the experience (Eslami et al., 2015; Kizilcec, 2016), and the findings of a large scale study in which participants from all over the world expressed concerns about the potential negative impact of algorithms (Newman et al., 2016). We speculate that participants perhaps don't see this kind of transparency information as a significant benefit to their experience. The potentially negative impact that presenting too much transparency information can have on the user experience is a concern that has been expressed by news media representatives (Diakopoulos and Koliska, 2016). In terms of personalization, participants indicated that the most important criterion was the ability to explicitly configure the system according to their own preferences. This indicates that personalization is a valued feature, as found by Chowdhury and Landoni (2006), but that users want to be their own editors and fully control their media diet, a finding mentioned by Newman et al. (2016).

Our results provide a clear hierarchy of importance of the criteria that users would like to manipulate in a news application managed by an algorithm, therefore answering the first research question. Based on these findings, and guidance from previous works, a working prototype of an algorithmic news aggregator was developed. It consists of a feed aggregator (Isbell, 2010) located in the lower left quadrant of the categorization presented by Coddington (2015) (Figure 2.1). The system was built on top of the MediaViz platform (Devezas et al., 2015), which provides a news archive comprising more than 4.3 million articles, and uses the open-source Lingo algorithm (Osiński et al., 2004) to select the stories. A user friendly UI, access to high quality and reputable sources, and personalization, which were some the highest rated features mentioned in Chowdhury and Landoni (2006), are among the functionalities provided. The UI layer is also responsible for conveying algorithmic transparency and control, using strategies such as signalling that the content is algorithmically processed, allow users to tweak their news feed by manipulating the algorithm's parameters, and provide textual descriptions of how it operates (Diakopoulos and Koliska, 2016).

The developed prototype incorporates some of the most relevant findings from this study and we believe it's a successful proof-of-concept of an algorithmic news application concerned with user control and transparency. It alleviates information overload by algorithmically condensing thousands of articles into groups of stories about a given day's most relevant topics, allows users to explicitly personalize their news feed – by selecting only the sources they want and manipulating the algorithm's parameters –, and presents algorithmic transparency information, namely signalling the algorithm's presence and describing how it operates. Therefore, the second research question presented was also responded.

Finally, it bears mention that this study entails some limitations, namely the usage of a non-probability sampling method which limits the generalization of the findings, and a questionnaire that was developed for this particular context, and therefore not validated. Nevertheless, we believe that our findings can be a useful reference for developers of algorithmic systems in the field of news media aiming to simultaneously provide user control, transparency, and an experience aligned with users' requirements.

6.1 Contributions

The contribution of this study is two-fold. First, it provides quantitative data collected from a considerable sample about the importance attributed to a broad set of criteria relevant in the context of news applications, from the users' perspective. Our belief about the quantitative findings' value is grounded in our own work. They were instrumental in informing the development of a functional prototype of an algorithmic news aggregator. The data also paves the way for news lines of research, including some that were identified in this document. They include determining the usage of news content criteria in stories produced by media organizations and shared by Portuguese users on social networks, and examining the degree to which they reflect the hierarchy of importance expressed by the participants. Identifying the reasons behind the moderate interest shown towards algorithmic information is another area worthy of examination. This work's second contribution is the developed prototype, which represents the embodiment of a set of strategies to avert information overload and provide user control and algorithmic transparency in an visual, interactive, and testable environment.

6.2 Future Work

The implementation of new features has already begun, particularly the inclusion of additional algorithm manipulation parameters. Future work includes the integration of a content analysis pipeline, which will be necessary to automatically classify the news in terms of their content, i.e., if they are good or bad news, if the people they mention are celebrities or from the power elite, or if they refer to relevant and impactful events or can be considered entertainment. The goal is to allow users to manipulate how much content with these criteria they would like in their news feed via UI controls. Other planned functionalities include the ability to turn off the algorithm and display an unfiltered news list, one strategy mentioned by Diakopoulos and Koliska (2016) which wasn't implemented, and the integration of more algorithms and the possibility to switch between them. Robust search functionality, a highly valued feature on news aggregators (Chowdhury and Landoni, 2006), is also part of our goals. Finally, and in tandem with the implementation of these features, we aim to gather information about the interaction with the application in order to assess how effective, efficient and satisfactory it is in terms of the issues it aims to address. Two strategies have been identified: anonymous data collection through usage logs, and observational tests with end-users.

References

Hunt Allcott and Matthew Gentzkow. Social Media and Fake News in the 2016 Election. Working Paper January, National Bureau of Economic Research, 2017.

C. W. Anderson. What aggregators do: Towards a networked concept of journalistic expertise in the digital age. *Journalism*, 14(8):1008–1023, nov 2013.

Association for Computing Machinery US Public Policy Council. Statement on algorithmic transparency and accountability. *USACM press releases*, (January 12), 2017.

Engin Bozdag. Bias in algorithmic filtering and personalization. *Ethics and Information Technology*, 15(3):209–227, 2013.

V.E. Bozdag. *Bursting the Filter Bubble: Democracy, Design, and Ethics*. PhD thesis, 2015.

Encyclopædia Britannica. Algorithm. URL https://www.britannica.com/topic/algorithm (Accessed: 2017-01-09).

Sudatta Chowdhury and Monica Landoni. News aggregator services: user expectations and experience. *Online Information Review*, 30(2):100–115, mar 2006.

Mark Coddington. *Telling Secondhand Stories: News Aggregation and the Production of Journalistic Knowledge*. PhD thesis, The University of Texas at Austin, 2015.

Tiago Devezas, Sérgio Nunes, and María Teresa Rodríguez. MediaViz: An Interactive Visualization Platform for Online Media Studies. In *Proceedings of the 2015 International Workshop on Human-centric Independent Computing - HIC '15*, pages 7–11, New York, New York, USA, 2015. ACM, ACM Press.

Nicholas Diakopoulos. Algorithmic Accountability: On the Investigation of Black Boxes. Technical report, Tow Center for Digital Journalism, 2014.

Nicholas Diakopoulos. Accountability in algorithmic decision making. *Communications of the ACM*, 59(2):56–62, jan 2016.

Nicholas Diakopoulos and Michael Koliska. Algorithmic Transparency in the News Media. *Digital Journalism*, 0811(July):1–20, jul 2016.

Nicholas Diakopoulos, Stephen Cass, and Joshua Romero. Data-Driven Rankings : The Design and Development of the IEEE Top Programming Languages News App. In *Proceedings of the Symposium on Computation+ Journalism*, pages 1–5, 2014.

Robert Epstein. How Google Could Rig the 2016 Election, 2015. URL http://www.politico.com/magazine/story/2015/08/how-google-could-rig-the-2016-election-121548 (Accessed: 2017-01-09).

REFERENCES

Robert Epstein and Ronald E Robertson. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences of the United States of America*, 112(33):E4512–21, aug 2015.

Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. "I always assumed that I wasn't really that close to [her]": Reasoning about Invisible Algorithms in News Feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*, pages 153–162, New York, New York, USA, 2015. ACM Press.

Facebook. How News Feed Works, 2016a. URL https://nonprofits.fb.com/2016/06/17/how-news-feed-works/ (Accessed: 2017-01-10).

Facebook. News Feed FYI: Addressing Hoaxes and Fake News, 2016b. URL http://newsroom.fb.com/news/2016/12/news-feed-fyi-addressing-hoaxes-and-fake-news/ (Accessed: 2017-01-25).

Batya Friedman and Helen Nissenbaum. Bias in computer systems. *ACM Transactions on Information Systems*, 14(3):330–347, jul 1996.

Kevin Hamilton, Karrie Karahalios, Christian Sandvig, and Motahhare Eslami. A Path to Understanding the Effects of Algorithm Awareness. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '14, pages 631–642, New York, NY, USA, 2014. ACM.

Tony Harcup and Deirdre O'Neill. What is news? *Journalism Studies*, 9699(March):1–19, mar 2016.

Kimberley Isbell. The Rise of the News Aggregator: Legal Implications and Best Practices. *The Berkman Center for Internet & Society at Harvard University*, 7641:28, 2010.

René F. Kizilcec. How Much Information? Effects of Transparency on Trust in an Algorithmic Interface. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, pages 2390–2395, 2016.

Nic Newman, David A. L. Levy, and Rasmus Kleis Nielsen. Reuters Institute Digital News Report 2015. Technical report, 2015.

Nic Newman, Richard Fletcher, David A. L. Levy, and Rasmus Kleis Nielsen. Reuters Institute Digital News Report 2016. Technical report, Reuters Institute for the Study of Journalism, Oxford University, 2016.

Publicações Obercom. Perfil sociodemográfico do consumo de notícias em Portugal. Technical report, Obercom, 2016.

Tim O'Reilly. What Is Web 2.0, 2005. URL http://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html?page=1 (Accessed: 2017-01-30).

Stanisław Osiński, Jerzy Stefanowski, and Dawid Weiss. Lingo: Search Results Clustering Algorithm Based on Singular Value Decomposition. In *Intelligent Information Processing and Web Mining*, pages 359–368. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

Hannah Jane Parkinson. Click and elect: how fake news helped Donald Trump win a real election, 2016. URL https://www.theguardian.com/commentisfree/2016/nov/14/fake-news-donald-trump-election-alt-right-social-media-tech-companies (Accessed: 2017-01-25).

REFERENCES

Pew Research Center. News Use Across Social Media Platforms 2016, 2016a. URL http://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/ (Accessed: 2017-01-28).

Pew Research Center. Digital News Audience: Fact Sheet, 2016b. URL http://www.journalism.org/2016/06/15/digital-news-audience-fact-sheet/ (Accessed: 2017-01-30).

Elia Powers. My News Feed is Filtered? *Digital Journalism*, 0811(February):1–21, feb 2017.

Emilee Rader. Examining user surprise as a symptom of algorithmic filtering. *International Journal of Human-Computer Studies*, 98:72–88, 2017.

Craig Silverman. This Analysis Shows How Viral Fake Election News Stories Outperformed Real News On Facebook, 2016. URL https://www.buzzfeed.com/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook?utm{_}term=.tmrrLYOyp{#}.muEkK8oG1 (Accessed: 2017-01-26).

James Stanyer. Web 2.0 and the Transformation of News and Journalism. *Routledge handbook of Internet politics*, pages 201–213, 2009.

Manuel Wendelin, Ines Engelmann, and Julia Neubarth. User Rankings and Journalistic News Selection. *Journalism Studies*, 18(2):135–153, 2017.

Appendix A

Questionnaire

A.1 English version

News criteria and computer-selected news

Suppose there is an application that has access to the news published by all Portuguese online news media. This application is able to condense and present, in a single page, the most relevant news according to a series of criteria. In a first phase, the list of news presented is selected by a computer program. Users of the application can then manipulate the criteria used by the program and change the news list according to their preferences.

The main objective of this questionnaire is to understand the most important criteria for you if you could create your own online journal. The questionnaire is divided into three main parts. The first part is related to the individual habits of consumption of news. The second aims to assess the subjective importance of the criteria regarding the content and diversity of the news, and the sources that produce it. The third part refers to the computer program that selects the news presented.

We are only interested in your opinion, so there are no good or bad answers. Your answers are strictly anonymous and confidential. Thanks for the collaboration.

*Obrigatório

1. **Age** *

   Marcar apenas uma oval.

   - ◯ 15-24
   - ◯ 25-34
   - ◯ 35-44
   - ◯ 45-54
   - ◯ 55-64
   - ◯ 65+
   - ◯ Won't answer

2. **Gender** *

   Marcar apenas uma oval.

   - ◯ Male
   - ◯ Female
   - ◯ Other
   - ◯ Won't answer

3. **Occupation** *

   If you are a student, please indicate the academic degree (e.g. undergraduate/master's/doctoral student)

   _____

4. **Scientific area of training** *
   Your main scientific area of training (eg, computer engineering, information science, communication sciences, psychology, etc.)

5. **Academic degree** *
   The highest academic degree you have obtained so far.

News consumption habits

6. **Typically, how often do you consume news?** *
   Marcar apenas uma oval.

   - ◯ Multiple times per hour
   - ◯ Once per hour
   - ◯ Several times a day
   - ◯ Once a day
   - ◯ Several times a week
   - ◯ Once a week
   - ◯ 2-3 times per month
   - ◯ Once a month
   - ◯ Less than once a month
   - ◯ Never

7. **What is your interest in news?** *
   Marcar apenas uma oval.

   - ◯ Very interested
   - ◯ Interested
   - ◯ Moderately interested
   - ◯ Little interested
   - ◯ Not at all interested

8. **Which media do you use to consume news? (Choose all that apply)** *
Marcar tudo o que for aplicável.

- [ ] Print media (newspapers, magazines, etc.)
- [ ] Radio
- [ ] Television
- [ ] The media's websites/applications
- [ ] Websites/applications that aggregate content from various media
- [ ] Social networks
- [ ] Blogs
- [ ] None of these
- [ ] Outra: _____

News criteria

Indicate the degree of importance of each of the following criteria in your perspective as a news consumer according to the scale presented.

Regarding the content of the presented news

9. **Timeliness (news published on top of the moment)** *
Marcar apenas uma oval por linha.

	Very important	Important	Moderately important	Little important	Not at all important	I don't know/Won't answer
	◯	◯	◯	◯	◯	◯

10. **Proximity (news about events near where I am)** *
Marcar apenas uma oval por linha.

	Very important	Important	Moderately important	Little important	Not at all important	I don’t know/Won’t answer
	◯	◯	◯	◯	◯	◯

11. **Conflict (news about controversies, strikes, wars)** *
Marcar apenas uma oval por linha.

	Very important	Important	Moderately important	Little important	Not at all important	I don’t know/Won’t answer
	◯	◯	◯	◯	◯	◯

12. **Serendipity (discovering interesting information unexpectedly)** *
Marcar apenas uma oval por linha.

	Very important	Important	Moderately important	Little important	Not at all important	I don't know/Won't answer
	⬭	⬭	⬭	⬭	⬭	⬭

13. **Good news (news with positive tones - cures, scientific advances, recoveries, victories and celebrations)** *
Marcar apenas uma oval por linha.

	Very important	Important	Moderately important	Little important	Not at all important	I don't know/Won't answer
	⬭	⬭	⬭	⬭	⬭	⬭

14. **Sharing potential (news potentially able of generating shares and comments on Facebook, Twitter, and other social networks)** *
Marcar apenas uma oval por linha.

	Very important	Important	Moderately important	Little important	Not at all important	I don't know/Won't answer
	⬭	⬭	⬭	⬭	⬭	⬭

15. **Drama (news about a developing drama such as escapes, accidents, searches, rescues, court cases)** *
Marcar apenas uma oval por linha.

	Very important	Important	Moderately important	Little important	Not at all important	I don't know/Won't answer
	⬭	⬭	⬭	⬭	⬭	⬭

16. **Magnitude (important news due to the large number of people involved or potential impact)** *
Marcar apenas uma oval por linha.

	Very important	Important	Moderately important	Little important	Not at all important	I don't know/Won't answer
	⬭	⬭	⬭	⬭	⬭	⬭

Regarding the content of the presented news (cont.)

17. **Relevance (news about groups or nations understood as influential or historically and culturally similar to the audience)** *

Marcar apenas uma oval por linha.

	Very important	Important	Moderately important	Little important	Not at all important	I don't know/Won't answer
	◯	◯	◯	◯	◯	◯

18. **Surprise (news with an element of surprise, contrast, out of the ordinary)** *

Marcar apenas uma oval por linha.

	Very important	Important	Moderately important	Little important	Not at all important	I don't know/Won't answer
	◯	◯	◯	◯	◯	◯

19. **Power elite (news about powerful entities such as individuals, organizations, institutions or companies)** *

Marcar apenas uma oval por linha.

	Very important	Important	Moderately important	Little important	Not at all important	I don't know/Won't answer
	◯	◯	◯	◯	◯	◯

20. **Audio-visual (accompanying news from photos, videos, audio and interesting graphics)** *

Marcar apenas uma oval por linha.

	Very important	Important	Moderately important	Little important	Not at all important	I don't know/Won't answer
	◯	◯	◯	◯	◯	◯

21. **Entertainment (light news - sex, sport, show business, animals, humorous treatment)** *

Marcar apenas uma oval por linha.

	Very important	Important	Moderately important	Little important	Not at all important	I don't know/Won't answer
	◯	◯	◯	◯	◯	◯

22. **Celebrities (news about famous people)** *

Marcar apenas uma oval por linha.

	Very important	Important	Moderately important	Little important	Not at all important	I don't know/Won't answer
	◯	◯	◯	◯	◯	◯

23. **Exclusivity (original news and first published in a source)** *
Marcar apenas uma oval por linha.

	Very important	Important	Moderately important	Little important	Not at all important	I don't know/Won't answer

24. **Bad news (news with negative tones - deaths, injuries, defeats)** *
Marcar apenas uma oval por linha.

	Very important	Important	Moderately important	Little important	Not at all important	I don't know/Won't answer

25. **Follow-up (news about subjects already present in the news)** *
Marcar apenas uma oval por linha.

	Very important	Important	Moderately important	Little important	Not at all important	I don't know/Won't answer

26. **Another**
If there are criteria that are important to you and are not mentioned above, please indicate them here.

Regarding the diversity of perspectives in the presented news

Indicate the degree of importance of each of the following criteria in your perspective as a news consumer according to the scale presented.

27. **News that are in line with my perspectives and views** *
Marcar apenas uma oval por linha.

	Very important	Important	Moderately important	Little important	Not at all important	I don't know/Won't answer

28. **News that present different perspectives and points of view from mine** *
Marcar apenas uma oval por linha.

	Very important	Important	Moderately important	Little important	Not at all important	I don't know/Won't answer

Regarding the sources that produce the news

29. **Journalistic prestige of the source** *
Marcar apenas uma oval por linha.

	Very important	Important	Moderately important	Little important	Not at all important	I don't know/Won't answer

30. **Size of the source operation (the number of people working there)** *
Marcar apenas uma oval por linha.

	Very important	Important	Moderately important	Little important	Not at all important	I don't know/Won't answer

31. **Professionalism of the source (if the news are written by professional journalists)** *
Marcar apenas uma oval por linha.

	Very important	Important	Moderately important	Little important	Not at all important	I don't know/Won't answer

32. **Reach of the source (if it reaches many people)** *
Marcar apenas uma oval por linha.

	Very important	Important	Moderately important	Little important	Not at all important	I don't know/Won't answer

33. **Another**
If there are criteria that are important to you and are not mentioned above, please indicate them here.

News selected by a computer program

Indicate the degree of importance of each of the following criteria in your perspective as a user of a news application selected by a computer program according to the scale presented.

Regarding the importance of information about the operation of a news application selected by a computer program

34. **Indication that news selection is done by a computer program** *

Marcar apenas uma oval por linha.

	Very important	Important	Moderately important	Little important	Not at all important	I don't know/Won't answer
	◯	◯	◯	◯	◯	◯

35. **Information about the various steps used by the computer program to select the news presented** *

Marcar apenas uma oval por linha.

	Very important	Important	Moderately important	Little important	Not at all important	I don't know/Won't answer
	◯	◯	◯	◯	◯	◯

Regarding the degree of personalization of the news list

In this context, personalization means adapting the news list to the specific profile of each user.

36. **Without personalization (I do not want the program to create a news list adapted to my profile)** *

Marcar apenas uma oval por linha.

	Very important	Important	Moderately important	Little important	Not at all important	I don't know/Won't answer
	◯	◯	◯	◯	◯	◯

37. **Explicit personalization (I want to be able to configure the program that generates the news list according to my criteria)** *

Marcar apenas uma oval por linha.

	Very important	Important	Moderately important	Little important	Not at all important	I don't know/Won't answer
	◯	◯	◯	◯	◯	◯

38. **Implicit personalization based on my actions (I want the program to display news based on my action history - searches, clicks, comments, 'likes', etc.)** *

Marcar apenas uma oval por linha.

	Very important	Important	Moderately important	Little important	Not at all important	I don't know/Won't answer
	◯	◯	◯	◯	◯	◯

39. **Implicit personalization based on my friends actions (I want the program to display news based on my friends' action history - searches, clicks, comments, likes, etc.)** *

Marcar apenas uma oval por linha.

	Very important	Important	Moderately important	Little important	Not at all important	I don't know/Won't answer
	◯	◯	◯	◯	◯	◯

40. **Implicit community-based personalization (I want the program to display news based on the action history of most users - searches, clicks, comments, 'likes', etc.)** *

Marcar apenas uma oval por linha.

	Very important	Important	Moderately important	Little important	Not at all important	I don't know/Won't answer
	◯	◯	◯	◯	◯	◯

41. **Comments**

If you want to leave a comment, please use this space.

A.2 Portuguese version

Critérios noticiosos e notícias seleccionadas por computador

Suponha que existe uma aplicação que tem acesso às notícias publicadas por todos os meios noticiosos online portugueses. Esta aplicação é capaz de condensar e apresentar, numa única página, as notícias mais relevantes de acordo com uma série de critérios. Numa primeira fase, a lista de notícias apresentadas é seleccionada por um programa de computador. Os utilizadores da aplicação podem então manipular os critérios usados pelo programa e alterar a lista de notícias de acordo com as suas preferências.

O principal objectivo deste questionário é perceber quais os critérios mais importantes para si caso tivesse a possibilidade de criar o seu próprio jornal online. O questionário está dividido em três partes principais. A primeira parte está relacionada com os hábitos individuais de consumo de notícias. A segunda visa aferir a importância subjectiva dos critérios relativos ao conteúdo e diversidade das notícias, e às fontes que as produzem. A terceira parte refere-se ao programa de computador que selecciona as notícias apresentadas.

Apenas estamos interessados na sua opinião, pelo que não há respostas boas nem más. As suas respostas são estritamente anónimas e confidenciais. Obrigado pela colaboração.

*Obrigatório

1. **Idade** *
   Marcar apenas uma oval.

   - ◯ 15-24
   - ◯ 25-34
   - ◯ 35-44
   - ◯ 45-54
   - ◯ 55-64
   - ◯ 65+
   - ◯ Não respondo

2. **Sexo** *
   Marcar apenas uma oval.

   - ◯ Masculino
   - ◯ Feminino
   - ◯ Outro
   - ◯ Não respondo

3. **Profissão** *
   Se for estudante, por favor indique o grau académico (p.e. estudante de licenciatura/mestrado/doutoramento))

   _____

4. **Área científica de formação** *
   A área científica principal da sua formação (p.e., engenharia informática, ciência da informação, ciências da comunicação, psicologia, etc.)

5. **Grau académico** *
   Qual o grau académico mais elevado que obteve até ao momento.

Hábitos de consumo de notícias

6. **Tipicamente, com que regularidade consome notícias?** *
   Marcar apenas uma oval.

   - ◯ Várias vezes por hora
   - ◯ Uma vez por hora
   - ◯ Várias vezes por dia
   - ◯ Uma vez por dia
   - ◯ Várias vezes por semana
   - ◯ Uma vez por semana
   - ◯ 2-3 vezes por mês
   - ◯ Uma vez por mês
   - ◯ Menos de uma vez por mês
   - ◯ Nunca

7. **Qual o seu interesse por notícias?** *
   Marcar apenas uma oval.

   - ◯ Muito interessado
   - ◯ Interessado
   - ◯ Moderadamente interessado
   - ◯ Pouco interessado
   - ◯ Nada interessado

8. **Que meios usa preferencialmente para consumir notícias? (Escolha todas as que se aplicam)** *
   Marcar tudo o que for aplicável.

   - [ ] Meios impressos (jornais, revistas, etc.)
   - [ ] Rádio
   - [ ] Televisão
   - [ ] Websites/aplicações dos meios de comunicação
   - [ ] Websites/aplicações que agregam conteúdos de vários meios de comunicação
   - [ ] Redes sociais
   - [ ] Blogues
   - [ ] Nenhum destes
   - [ ] Outra: _____

Critérios noticiosos

Indique o grau de importância de cada um dos seguintes critérios na sua perspectiva como consumidor/a de notícias de acordo com a escala apresentada.

Quanto ao conteúdo das notícias apresentadas

9. **Actualidade (notícias publicadas em cima do momento)** *
   Marcar apenas uma oval por linha.

	Muito importante	Importante	Moderadamente importante	Pouco importante	Nada importante	Não sei/Não respondo
	◯	◯	◯	◯	◯	◯

10. **Proximidade (notícias sobre eventos próximos de onde me encontro)** *
    Marcar apenas uma oval por linha.

	Muito importante	Importante	Moderadamente importante	Pouco importante	Nada importante	Não sei/Não respondo
	◯	◯	◯	◯	◯	◯

11. **Conflito (notícias sobre controvérsias, greves, guerras)** *
    Marcar apenas uma oval por linha.

	Muito importante	Importante	Moderadamente importante	Pouco importante	Nada importante	Não sei/Não respondo
	◯	◯	◯	◯	◯	◯

12. **Serendipidade (descoberta de informação interessante de forma inesperada)** *
Marcar apenas uma oval por linha.

	Muito importante	Importante	Moderadamente importante	Pouco importante	Nada importante	Não sei/Não respondo
	◯	◯	◯	◯	◯	◯

13. **Boas notícias (notícias com tons positivos - curas, avanços científicos, recuperações, vitórias e celebrações)** *
Marcar apenas uma oval por linha.

	Muito importante	Importante	Moderadamente importante	Pouco importante	Nada importante	Não sei/Não respondo
	◯	◯	◯	◯	◯	◯

14. **Potencial de partilha (notícias potencialmente capazes de gerar partilhas e comentários no Facebook, Twitter e outras redes sociais)** *
Marcar apenas uma oval por linha.

	Muito importante	Importante	Moderadamente importante	Pouco importante	Nada importante	Não sei/Não respondo
	◯	◯	◯	◯	◯	◯

15. **Drama (notícias sobre um drama em desenvolvimento como fugas, acidentes, buscas, salvamentos, casos judiciais)** *
Marcar apenas uma oval por linha.

	Muito importante	Importante	Moderadamente importante	Pouco importante	Nada importante	Não sei/Não respondo
	◯	◯	◯	◯	◯	◯

16. **Magnitude (notícias importantes devido ao grande número de pessoas envolvidas ou impacto potencial)** *
Marcar apenas uma oval por linha.

	Muito importante	Importante	Moderadamente importante	Pouco importante	Nada importante	Não sei/Não respondo
	◯	◯	◯	◯	◯	◯

Quanto ao conteúdo das notícias apresentadas (cont.)

17. **Relevância (notícias sobre grupos ou nações entendidos como influentes ou historica e culturalmente similares à audiência)** *

Marcar apenas uma oval por linha.

	Muito importante	Importante	Moderadamente importante	Pouco importante	Nada importante	Não sei/Não respondo
	◯	◯	◯	◯	◯	◯

18. **Surpresa (notícias com um elemento de surpresa, contraste, fora do comum)** *

Marcar apenas uma oval por linha.

	Muito importante	Importante	Moderadamente importante	Pouco importante	Nada importante	Não sei/Não respondo
	◯	◯	◯	◯	◯	◯

19. **Elite do poder (notícias sobre entidades poderosas como indivíduos, organizações, instituições ou empresas)** *

Marcar apenas uma oval por linha.

	Muito importante	Importante	Moderadamente importante	Pouco importante	Nada importante	Não sei/Não respondo
	◯	◯	◯	◯	◯	◯

20. **Áudio-visual (notícias acompanhas de fotos, vídeos, áudio e gráficos interessantes)** *

Marcar apenas uma oval por linha.

	Muito importante	Importante	Moderadamente importante	Pouco importante	Nada importante	Não sei/Não respondo
	◯	◯	◯	◯	◯	◯

21. **Entretenimento (notícias ligeiras - sexo, desporto, mundo do espectáculo, animais, tratamento humorístico)** *

Marcar apenas uma oval por linha.

	Muito importante	Importante	Moderadamente importante	Pouco importante	Nada importante	Não sei/Não respondo
	◯	◯	◯	◯	◯	◯

22. **Celebridades (notícias sobre pessoas famosas)** *

Marcar apenas uma oval por linha.

	Muito importante	Importante	Moderadamente importante	Pouco importante	Nada importante	Não sei/Não respondo
	◯	◯	◯	◯	◯	◯

23. **Exclusividade (notícias originais e publicadas primeiro numa fonte)** *
Marcar apenas uma oval por linha.

	Muito importante	Importante	Moderadamente importante	Pouco importante	Nada importante	Não sei/Não respondo
	◯	◯	◯	◯	◯	◯

24. **Más notícias (notícias com tons negativos - mortes, feridos, derrotas)** *
Marcar apenas uma oval por linha.

	Muito importante	Importante	Moderadamente importante	Pouco importante	Nada importante	Não sei/Não respondo
	◯	◯	◯	◯	◯	◯

25. **Acompanhamento/'follow up' (notícias sobre assuntos já presentes nas notícias)** *
Marcar apenas uma oval por linha.

	Muito importante	Importante	Moderadamente importante	Pouco importante	Nada importante	Não sei/Não respondo
	◯	◯	◯	◯	◯	◯

26. **Outro**
Se houver critérios que são importantes para si e não estão mencionados acima, por favor indique-os aqui.

Quanto à diversidade de perspectivas nas notícias apresentadas

Indique o grau de importância de cada um dos seguintes critérios na sua perspectiva como consumidor/a de notícias de acordo com a escala apresentada.

27. **Notícias que estão de acordo com as minhas perspectivas e pontos de vista** *
Marcar apenas uma oval por linha.

	Muito importante	Importante	Moderadamente importante	Pouco importante	Nada importante	Não sei/Não respondo
	◯	◯	◯	◯	◯	◯

28. **Notícias que apresentam perspectivas e pontos de vista diferentes dos meus** *
Marcar apenas uma oval por linha.

	Muito importante	Importante	Moderadamente importante	Pouco importante	Nada importante	Não sei/Não respondo
	◯	◯	◯	◯	◯	◯

Quanto às fontes que produzem as notícias

29. **Prestígio jornalístico da fonte** *
Marcar apenas uma oval por linha.

	Muito importante	Importante	Moderadamente importante	Pouco importante	Nada importante	Não sei/Não respondo
	◯	◯	◯	◯	◯	◯

30. **Dimensão da operação da fonte (o número de pessoas que lá trabalham)** *
Marcar apenas uma oval por linha.

	Muito importante	Importante	Moderadamente importante	Pouco importante	Nada importante	Não sei/Não respondo
	◯	◯	◯	◯	◯	◯

31. **Profissionalismo da fonte (se as notícias são escritas por jornalistas profissionais)** *
Marcar apenas uma oval por linha.

	Muito importante	Importante	Moderadamente importante	Pouco importante	Nada importante	Não sei/Não respondo
	◯	◯	◯	◯	◯	◯

32. **Alcance da fonte (se chega a muitas pessoas)** *
Marcar apenas uma oval por linha.

	Muito importante	Importante	Moderadamente importante	Pouco importante	Nada importante	Não sei/Não respondo
	◯	◯	◯	◯	◯	◯

33. **Outro**
Se houver critérios que são importantes para si e não estão mencionados acima, por favor indique-os aqui.

Notícias seleccionadas por um programa de computador

Indique o grau de importância de cada um dos seguintes critérios na sua perspectiva como utilizador/a de uma aplicação de notícias seleccionadas por um programa de computador de acordo com a escala apresentada.

Quanto à importância da informação sobre o funcionamento de uma aplicação de notícias seleccionadas por um programa de computador

34. **Indicação de que a selecção das notícias é feita por um programa de computador** *

Marcar apenas uma oval por linha.

	Muito importante	Importante	Moderadamente importante	Pouco importante	Nada importante	Não sei/Não respondo
	◯	◯	◯	◯	◯	◯

35. **Informação sobre os vários passos usados pelo programa de computador para seleccionar as notícias apresentadas** *

Marcar apenas uma oval por linha.

	Muito importante	Importante	Moderadamente importante	Pouco importante	Nada importante	Não sei/Não respondo
	◯	◯	◯	◯	◯	◯

Quanto ao grau de personalização da lista de notícias

Neste contexto, personalização significa adaptar a lista de notícias ao perfil específico de cada utilizador.

36. **Sem personalização (não quero que o programa crie uma lista de notícias adaptadas ao meu perfil)** *

Marcar apenas uma oval por linha.

	Muito importante	Importante	Moderadamente importante	Pouco importante	Nada importante	Não sei/Não respondo
	◯	◯	◯	◯	◯	◯

37. **Personalização explícita (quero poder configurar o programa que gera a lista de notícias de acordo com os meus critérios)** *

Marcar apenas uma oval por linha.

	Muito importante	Importante	Moderadamente importante	Pouco importante	Nada importante	Não sei/Não respondo
	◯	◯	◯	◯	◯	◯

38. **Personalização implícita baseada nas minhas acções (quero que o programa exiba notícias com base no meu histórico de acções - pesquisas, cliques, comentários, 'gostos', etc.)** *

Marcar apenas uma oval por linha.

	Muito importante	Importante	Moderadamente importante	Pouco importante	Nada importante	Não sei/Não respondo
	◯	◯	◯	◯	◯	◯

39. **Personalização implícita baseada nas acções dos meu amigos (quero que o programa exiba notícias com base no histórico de acções dos meus amigos - pesquisas, cliques, comentários, 'gostos', etc.)** *

Marcar apenas uma oval por linha.

	Muito importante	Importante	Moderadamente importante	Pouco importante	Nada importante	Não sei/Não respondo
	◯	◯	◯	◯	◯	◯

40. **Personalização implícita baseada nas acções da comunidade (quero que o programa exiba notícias com base no histórico de acções da maioria dos utilizadores - pesquisas, cliques, comentários, 'gostos', etc.)** *

Marcar apenas uma oval por linha.

	Muito importante	Importante	Moderadamente importante	Pouco importante	Nada importante	Não sei/Não respondo
	◯	◯	◯	◯	◯	◯

41. **Comentários**

Se quiser deixar algum comentário, por favor utilize este espaço.

Appendix B

Email sent to the participants

B.1 English version

Good evening,

In the context of a dissertation from the Master of Multimedia at UP, we aim to evaluate the subjective importance of a set o criteria used to create a newspaper tailored for each user.

In that regard, we ask for your collaboration in filling a questionnaire (approximately 5 minutes), whose link can be found below.

https://goo.gl/forms/lzuc7iNIcFO70zY62

Thank you in advance,

B.2 Portuguese version

Boa noite,

No âmbito de uma dissertação do Mestrado em Multimédia da UP, pretende-se avaliar a importância subjectiva de um conjunto de critérios usados para criar um jornal à medida de cada utilizador.

Nesse sentido, solicita-se a sua colaboração no preenchimento de um questionário (aproximadamente 5 minutos), cujo link se encontra abaixo.

https://goo.gl/forms/lzuc7iNIcFO70zY62

Desde já obrigado